

VISUAL TRANSFER LEARNING USING KNOWLEDGE GRAPHS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF
COMPUTATIONAL LINGUISTICS AT THE FACULTY II
OF TRIER UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Sebastian Monka

March 2023

© Copyright by Sebastian Monka 2024
All Rights Reserved

1. Gutachter: *Prof. Dr. Achim Rettinger*
2. Gutachter: *Prof. Dr. Volker Tresp*

Abstract

While humans find it easy to process visual information from the real world, machines struggle with this task due to the unstructured and complex nature of the information. *Computer vision* (CV) is the approach of artificial intelligence that attempts to automatically analyze, interpret, and extract such information. Recent CV approaches mainly use *deep learning* (DL) due to its very high accuracy. DL extracts useful features from unstructured images in a training dataset to use them for specific real-world tasks. However, DL requires a large number of parameters, computational power, and meaningful training data, which can be noisy, sparse, and incomplete for specific domains. Furthermore, DL tends to learn correlations from the training data that do not occur in reality, making DNNs poorly generalizable and error-prone.

Therefore, the field of visual transfer learning is seeking methods that are less dependent on training data and are thus more applicable in the constantly changing world. One idea is to enrich DL with prior knowledge. *Knowledge graphs* (KG) serve as a powerful tool for this purpose because they can formalize and organize prior knowledge based on an underlying ontological schema. They contain symbolic operations such as logic, rules, and reasoning, and can be created, adapted, and interpreted by domain experts. Due to the abstraction potential of symbols, KGs provide good prerequisites for generalizing their knowledge. To take advantage of the generalization properties of KG and the ability of DL to learn from large-scale unstructured data, attempts have long been made to combine explicit graph and implicit vector representations. However, with the recent development of knowledge graph embedding methods, where a graph is transferred into a vector space, new perspectives for a combination in vector space are opening up.

In this work, we attempt to combine prior knowledge from a KG with DL to improve visual transfer learning using the following steps: First, we explore the potential benefits of using prior knowledge encoded in a KG for DL-based visual transfer learning. Second, we investigate approaches that already combine KG and DL and create a categorization based on their general idea of knowledge integration. Third, we propose a novel method for the specific category of using the *knowledge graph as a trainer*, where a DNN is trained to adapt to a representation given by prior knowledge of a KG. Fourth, we extend the proposed method by extracting relevant context in the form of a subgraph of the KG to investigate the relationship between prior knowledge and performance on a specific CV

task. In summary, this work provides deep insights into the combination of KG and DL, with the goal of making DL approaches more generalizable, more efficient, and more interpretable through prior knowledge.

Zusammenfassung

Während es Menschen leicht fällt, visuelle Informationen der realen Welt zu verarbeiten, gestaltet sich dies bei Maschinen deutlich schwieriger, da die Informationen unstrukturiert und komplex sind. *Computer Vision* (CV) ist der Ansatz der künstlichen Intelligenz, der versucht solche Informationen automatisch zu analysieren, zu interpretieren und zu extrahieren. Neuere CV-Ansätze verwenden aufgrund der sehr hohen Genauigkeit hauptsächlich *Deep Learning* (DL). Dabei werden nützliche Merkmale aus unstrukturierten Bildern eines Trainingsdatensatzes extrahiert, um diese für bestimmte Aufgaben in der realen Welt zu nutzen. DL erfordert jedoch eine große Menge an Parametern, Rechenleistung und aussagekräftigen Trainingsdaten, die für bestimmte Bereiche verrauscht, spärlich und unvollständig sein können. Darüber hinaus neigt DL dazu, aus den Trainingsdaten auch Korrelationen zu lernen, die in der Realität nicht vorkommen. Diese Eigenschaften machen DL-Modelle schlecht generalisierbar und fehleranfällig.

Deshalb sucht das Forschungsfeld des *Visual Transfer Learning* nach Methoden, die weniger stark von den Trainingsdaten abhängen und somit besser in der sich ständig verändernden Welt anwendbar sind. Eine dieser Methoden versucht, DL mit Vorwissen anzureichern. Dazu dienen *Knowledge Graphs* (KG) als leistungsfähiges Werkzeug, da sie Vorwissen auf Basis eines zugrunde liegenden ontologischen Schemas gut formalisieren und organisieren können. Sie beinhalten symbolische Operationen wie Logik, Regeln und Schlussfolgerungen und können von Domänenexperten erstellt, angepasst und interpretiert werden. Aufgrund des Abstraktionspotentials von Symbolen bieten KGs gute Voraussetzungen, ihr Wissen zu generalisieren. Um die Vorteile von KG und DL zu vereinen, wurde schon lange Zeit versucht, explizite Graph- und implizite Vektorrepräsentationen zu kombinieren. Durch die Entwicklung von *Knowledge Graph Embedding Methods*, bei denen ein Graph in den Vektorraum transferiert wird, eröffnen sich neue Perspektiven für eine Kombination.

In dieser Arbeit untersuchen wir die Kombination von KG und DL, um *Visual Transfer Learning* zu verbessern, anhand der folgenden Schritte: Erstens untersuchen wir die potenziellen Vorteile der Verwendung von in einem KG kodiertem Vorwissen für DL-basiertes *Visual Transfer Learning*. Zweitens fassen wir Ansätze zusammen, die bereits KG und DL kombinieren, und erstellen eine Kategorisierung auf der Grundlage ihrer allgemeinen Idee der Wissensintegration. Drittens schlagen wir eine neuartige Methode für die spezielle Kategorie der Verwendung des Wissensgraphen als

Trainer vor, bei der ein *Deep Neural Network* (DNN) so trainiert wird, dass es sich an eine durch das Vorwissen eines KG gegebene Darstellung anpasst. Viertens erweitern wir die vorgeschlagene Methode durch die Extraktion von relevantem Kontext in Form eines Subgraphen des KG, um die Beziehung zwischen dem Vorwissen und der Leistung bei einer bestimmten CV-Aufgabe zu untersuchen. Zusammenfassend lässt sich sagen, dass diese Arbeit tiefe Einblicke in die Kombination von KG und DL bietet, mit dem Ziel, DL-Ansätze durch Vorwissen verallgemeinerbarer, effizienter und interpretierbarer zu machen.

Acknowledgements

I owe a debt of gratitude to numerous people for their invaluable assistance and support in making this dissertation possible. Foremost among them are my two advisors, Achim Rettinger and Lavdim Halilaj. Their guidance and encouragement have been instrumental in shaping my research direction and goals. Achim, you have taught me to trust my intuition and pursue my ideas with confidence. Thank you for always being available to offer your expertise and support whenever I needed it. Lavdim, your critical insights and thoughtful discussions have helped me navigate complex decisions and refine my work. I am grateful for the countless hours you have spent reading my drafts and providing feedback. I would also like to extend my deep appreciation to Volker Tresp for his flexibility, cooperation, and expertise, which have played a critical role in bringing this thesis to a successful conclusion.

I am grateful to Trier University, Faculty II, and the Department of Computational Linguistics, as well as Robert Bosch GmbH, Bosch Research, and the Bosch Center for Artificial Intelligence for providing a conducive working environment. Additionally, I want to thank the research project "KI Delta Learning" (project number: 19A19013D) funded by the Federal Ministry for Economic Affairs and Energy (BMWi) on the basis of a decision by the German Bundestag for its financial and organizational support.

I am deeply indebted to my family, Ulrike, Christian, Carolin, Julian, and Daniel, and my friends for their unwavering support, encouragement, and love. Their belief in me has been a source of strength and inspiration. I would also like to acknowledge my husky, Jibby, whose bright, cheeky, and loving nature has enriched my life. Last but not least, I would like to express my heartfelt gratitude to my partner, Selay. Your love, encouragement, and support have been the cornerstone of my success. We have overcome challenges and achieved milestones. Together, we master everything with ease.

Contents

Abstract	v
Zusammenfassung	vii
Acknowledgements	ix
1 Introduction	1
1.1 Motivation	2
1.2 Problem Definition and Challenges	3
1.3 Research Questions	6
1.4 Thesis Overview	8
1.4.1 Contributions	8
1.4.2 List of Publications	9
1.5 Outline of the Thesis	10
2 Preliminaries	13
2.1 Neuro-Symbolic AI	13
2.1.1 Human Intelligence	15
2.1.2 Artificial Intelligence	16
2.1.3 Combining Knowledge Graph and Deep Learning	17
2.2 Modalities of Data, Information, and Knowledge	18
2.2.1 Vision	18
2.2.2 Language	19
2.2.3 Knowledge Graph	20
2.3 Feature Extraction	22
2.3.1 Visual Features Extractor	23
2.3.2 Semantic Features Extractor	24
2.4 Inductive Biases in Deep Learning	27
2.4.1 Dataset	27

2.4.2	Data Augmentation	28
2.4.3	Labels	29
2.4.4	Network Architecture	30
2.4.5	Training Objective	30
2.4.6	Optimization Method	32
3	Transfer Learning using Prior Knowledge	35
3.1	Fundamentals of Transfer Learning	36
3.1.1	Transfer Learning Scenarios	37
3.1.2	Strategies in Transfer Learning	38
3.2	Domain Change in Deep Learning	39
3.2.1	Specialization-Generalization Trade-off in DL	40
3.2.2	Enabling Deep Learning to Generalize	41
3.3	Transfer Learning using Knowledge Graphs	42
3.3.1	Types of Prior Knowledge	43
3.3.2	Knowledge Graph as a Representation Format	44
3.4	Resources for Visual Transfer Learning using KG	45
3.4.1	Generic Knowledge Graphs	46
3.4.2	Image Datasets	48
3.5	Summary	54
4	Combinations of Knowledge Graphs and Deep Learning	55
4.1	Methodology	56
4.1.1	Literature Search	57
4.1.2	Literature Selection and Quality Assessment	57
4.2	A Categorization of Visual Transfer Learning using KG	58
4.2.1	Knowledge Graph as a Reviewer	60
4.2.2	Knowledge Graph as a Trainee	63
4.2.3	Knowledge Graph as a Trainer	65
4.2.4	Knowledge Graph as a Peer	68
4.3	Surveys Related to Visual Transfer Learning using KG	71
4.4	Summary	73
5	Learning Visual Models using a Knowledge Graph as a Trainer	75
5.1	Knowledge Graph as a Trainer	77
5.1.1	Knowledge Infusion	77
5.1.2	Adaptation to a Labeled Target Domain	78
5.2	Experiments	80

5.2.1	Scenario 1 - Domain Generalization	80
5.2.2	Scenario 2 - Supervised Domain Adaptation	84
5.3	Work Related to Learning Visual Models using a KG	86
5.4	Summary	88
6	Context-driven Visual Object Recognition based on Knowledge Graphs	89
6.1	Contextual Image Representations	91
6.2	Learning Contextual Image Representations	92
6.2.1	Contextual View Extraction	93
6.2.2	Contextual View Infusion	94
6.3	Experiments	96
6.3.1	Implementation details	96
6.3.2	Experiments on Cifar10	96
6.3.3	Experiments on Mini-ImageNet	99
6.4	Work Related to Contextual Visual Models based on KG	101
6.5	Discussion and Insights	102
6.6	Summary	103
7	Conclusions	105
7.1	Revisiting the Research Questions	106
7.2	Further Challenges and Open Issues	109
7.3	Future Work	111
	Bibliography	119

List of Figures

1.1	Visual transfer learning using KG	2
1.2	Challenges and research questions	4
2.1	Prerequisites for artificial intelligence	14
2.2	Modality of vision	19
2.3	Modality of language	20
2.4	DNN pipeline	22
2.5	Visual and semantic features extractor	23
2.6	Principle of a KGE method	25
2.7	Inductive biases in deep learning	28
3.1	Transfer learning scenarios	38
3.2	Strategies in transfer learning	39
3.3	Representation formats of KG	44
4.1	Categories of visual transfer learning using KG	59
4.2	Knowledge Graph as a Reviewer	61
4.3	Knowledge Graph as a Trainee	63
4.4	Knowledge Graph as a Trainer	66
4.5	Knowledge Graph as a Peer	68
5.1	Overview of the KG-NN approach	77
5.2	The KG-NN pipeline	79
5.3	Domain generalization results for KG-NN on Mini-ImageNet	82
5.4	The road sign knowledge graph	83
5.5	Domain generalization results for KG-NN on GTSRB	84
5.6	Domain adaptation results for KG-NN on GTSRB	85
6.1	Ambiguous figures	89
6.2	Framework to learn contextual DNNs	92

6.3	Types of context	93
6.4	Contextual view infusion	95
6.5	Qualitative evaluation for Cifar10	97
6.6	Qualitative evaluation for Mini-ImageNet	100
6.7	Evaluation of sample predictions for Mini-ImageNet	101

List of Tables

2.1	Graph models for KG	21
3.1	Overview of generic KGs	46
3.2	Datasets for visual transfer learning using prior knowledge	49
4.1	Categories of KG-DL combinations and their tasks	60
6.1	Quantitative evaluation for Cifar10	98
6.2	Quantitative evaluation for Mini-ImageNet	100

Chapter 1

Introduction

Artificial intelligence (AI) can be defined as the theory and development of computer systems capable of performing tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and language translation [1]. While the topic of AI used to be only part of research, it has recently attracted a lot of interest from general public. From self-driving cars to virtual assistants, AI is becoming an integral part of daily lives, and its impact is expected to continue to grow in the coming years. *Computer vision* (CV) is the research area of AI that deals with the visual perception of the environment and visual transfer learning is the subfield that focuses on improving the robustness of current CV approaches to naturally occurring visual changes.

On the one hand, *deep learning* (DL), as representative of *machine learning* (ML), has gained popularity in recent years due to its very high accuracy on CV tasks. Its main strength lies in the large-scale statistical exploitation of huge amounts of unstructured data and the availability of high computational power. For this purpose, DL models, i.e. *deep neural networks* (DNN), implicitly learn highly detailed features of images based on correlations in a training dataset and use them to recognize objects in the real world. On the other hand, symbolic AI was the main driver of AI before DL gained prominence. There was a general belief that symbols are the root of intelligent action and therefore symbols should be the main topic of AI [2].

In Figure 1.1, we provide an overview of visual transfer learning using *knowledge graphs* (KG), which combines prior knowledge of a KG with unstructured image data encoded by DL to improve the generalization of DL. KG, as an emerging technology of symbolic AI, provide a way to encode prior knowledge about entities, including objects, events, situations, or concepts, and their various relationships in a structured way. Due to their underlying ontological schema, they use symbolic operations such as logic, rules, and reasoning, and can be created, adapted, and interpreted by human experts. A KG as a large-scale semantic network of real-world entities is therefore well suited as additional and controlled input for unstructured image data. The research area of AI that combines symbolic AI and ML is referred to as *neuro-symbolic AI* (NSAI) [3] and includes the idea of combining prior knowledge of a KG with unstructured image data encoded by DL.

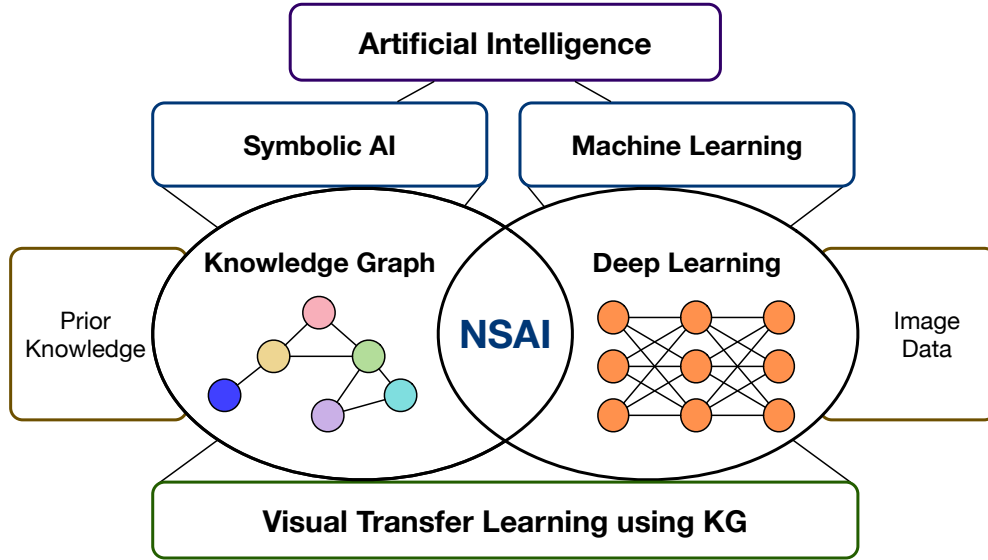


Figure 1.1: Visual transfer learning using KG is part of *neuro-symbolic AI* (NSAI) and combines prior knowledge with image data. While a *knowledge graph* (KG) as a representative of symbolic AI is well suited to store and process explicit prior knowledge, *deep learning* (DL) as a representative of machine learning can handle unstructured data, such as images, well.

1.1 Motivation

DL approaches achieve very high accuracy for a number of tasks and outperform traditional CV approaches by a large margin. Despite their widespread successes in AI-based applications like autonomous driving, they face a number of weaknesses. First, DL approaches are resource hungry. They require a large number of parameters, computational power, and expressive training data to perform well. In addition, the training data must be balanced, as well as carefully selected and labeled. Second, DL approaches have a poor ability to generalize. They tend to overfit the training data, learning spurious correlations which are not present when applying these models in the real world. Therefore, they require a training dataset that fully corresponds to the environment in which the model is used. If reality differs from the training data, DL can lead to unpredictable errors.

Visual transfer learning is the research area of CV with the goal of making DL approaches more robust to these failures and thus generalizable or transferable to related domains. Aiming to address the drawbacks of data-driven DL approaches, we consider another dimension and therefore propose to introduce prior knowledge about relevant correlations into the learning process. Prior knowledge of a domain, such as higher-level class relationships, should support DL to learn faster from the given training data but also to learn more robust features that generalize to related domains. Explicit prior knowledge is best encoded with symbolic approaches such as KGs. They can be constructed, customized, or interpreted by domain experts. Moreover, KGs are a promising technology to unify

prior knowledge about domains using ubiquitous concepts and an underlying ontological schema.

Therefore, we aim to combine a KG with DL to learn meaningful representations of large-scale unstructured image data using prior knowledge. However, the combination of symbolic, i.e. KG, and subsymbolic knowledge, i.e. DL, still poses a number of serious challenges, which will be discussed in the following.

1.2 Problem Definition and Challenges

DL and KG are coexisting technologies of AI that are commonly used for different problems due to their distinct strengths and weaknesses. DL, as part of ML, can make sense of highly unstructured data such as images or text using high-dimensional vector representations in a parallel manner. KGs, as part of symbolic AI, sequentially operate on symbol representations, use logic and rules, are explainable, and therefore well suited as an interface to human knowledge and experts. Due to their complementary strengths and weaknesses, the field of NSAI attempts to combine ML and symbolic AI, as illustrated in Figure 1.1.

However, the combination of explicit graph representations of symbolic AI and implicit vector representations of ML proves to be challenging. Therefore, most approaches to NSAI consider both fields independently. ML is responsible for extracting vector-based features from unstructured image data and transferring them into symbolic output representations, e.g. recognized entities. Symbolic AI then uses these extracted symbols to apply graph-based post-processing based on prior knowledge to improve the results. However, traditional NSAI approaches do not consider the fact that features learned by ML are highly dependent on the training data. If the domain changes, these features of the images are not recognized by DL, and symbolic post-processing no longer works. Therefore, to solve transfer learning problems other approaches need to be explored.

With the recent development of KGE methods, where a graph is transferred into a vector space, new possibilities have emerged to combine symbolic representations, e.g. KGs, with subsymbolic ones, e.g. DL. In the context of this work, we want to investigate the different ways of combining a KG with DL to improve visual transfer learning. We will refer to such combinations as KG-DL. Hereby, we face four main challenges (CH) and research questions (RQ), which are depicted in Figure 1.2 and introduced in the following.

Challenge 1 - Enhancing Transfer Learning using Prior Knowledge: Transfer learning is a powerful concept that involves the reuse of features from one domain to another. In DL, transfer learning is essential because models must learn distinct features from a training dataset and apply them to a test domain. However, for pure DL that relies solely on unstructured data, it is challenging to determine which features generalize and therefore are relevant for unknown domains. Moreover, there are hypotheses that assume that there can be no generalization without prior assumptions [4].

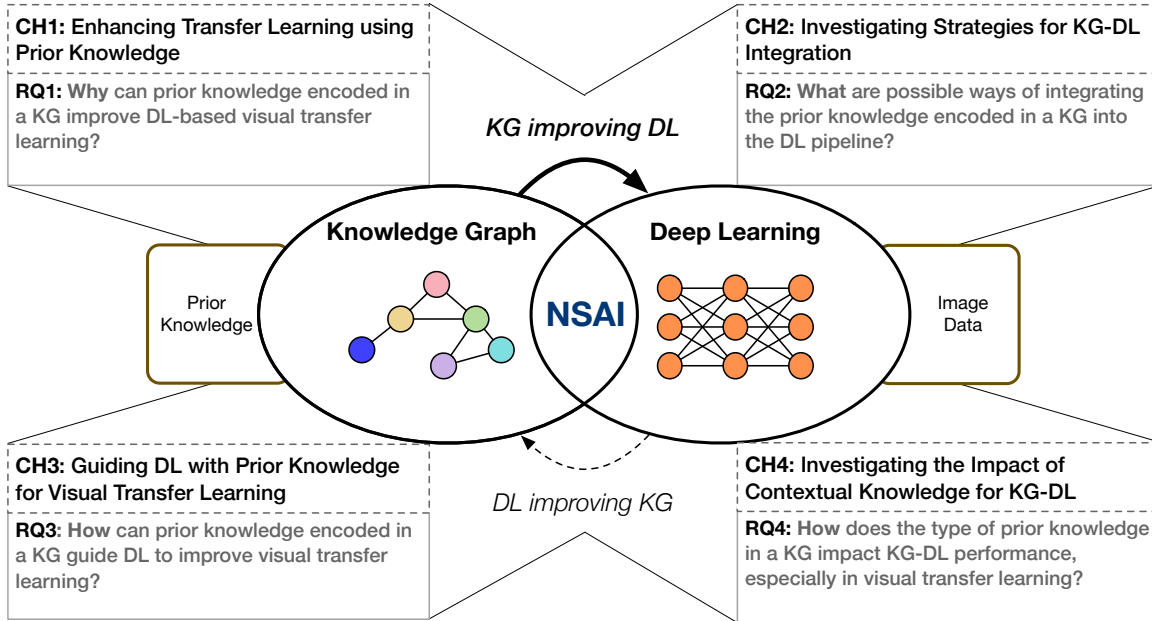


Figure 1.2: *Challenges (CH) and research questions (RQ)*. In this thesis, we identify four main challenges and research questions for using a *knowledge graph (KG)* to improve *deep learning (DL)* on visual transfer learning tasks.

In DL these prior assumptions are known as inductive biases and enable DNNs to learn transferable features for unseen domains. However, selecting the appropriate inductive bias for a given task remains an open question that requires human intuition and experience. Inductive biases can be considered a raw form of prior knowledge because they do not directly reflect human knowledge about a domain. Since, prior knowledge can be defined as additional information about the world, a natural question is if other types of prior knowledge could also be used to enhance DL-based transfer learning. Other types of prior knowledge include for instance additional measured data, data from other modalities, or explicit human expert knowledge. Therefore, the main challenge for transfer learning is to find other types of prior knowledge and suitable representation formats to enhance the transfer learning capabilities of DL.

Challenge 2 - Investigating Strategies for KG-DL Integration: Especially in the field of CV, DL approaches dominate the state of the art due to their high accuracy if trained on large-scale data. However, these methods tend to learn spurious correlations of the specific training data and face serious challenges if the domain changes. KGs can store prior knowledge, such as additional context and background information, that can be useful for the DL method to understand and interpret visual data correctly. KG-DL combinations have the potential to improve the performance of DNNs by using prior knowledge provided by a KG to learn more generalizable CV models. There

have been several approaches that attempt to incorporate prior knowledge into data-driven methods. However, it is still not clear what kind of combination leads to the best results or is best suited for a specific CV task.

The main challenge here is to understand the basic principles of KG-DL integration methods, as well as their strengths and weaknesses. Therefore, the methods need to be examined in terms of, for instance, how knowledge infusion is performed, in which transfer learning scenarios the methods are mainly used, and what impact KG has on DL. The specific characteristics of each integration method need to be explored to better understand how knowledge sharing between KG and DL occurs and how it enhances the final visual transfer learning task.

Challenge 3 - Guiding DL with Prior Knowledge for Visual Transfer Learning: While previous NSAI methods have mostly used KG and DL independently, by using DL for extracting features and KG for performing reasoning over the extracted features, there is a growing need to integrate the two methods more deeply. In particular, for visual transfer learning tasks where the application domain differs from the training domain, traditional NSAI methods fail similarly to DL-only approaches because relevant features cannot be extracted and reasoning cannot be performed. If a KG should improve the performance of DL on these tasks, prior knowledge must already be used while training the DNN to guide DL to learn robust features from data. With the development of KGE methods that transform a KG into vector space, new possibilities for interaction with vector-based DL have emerged and are worth exploring.

We thus hypothesize that deep integration of a KG’s prior knowledge when learning the DNN can significantly improve the generalizability of CV models. The main idea is to provide the DNN during training with information about what distinguishes a relevant correlation in the image data from a spurious correlation. The main challenge here is to develop and investigate a specific method that integrates the prior knowledge of the KG into the DNN at training time to better learn generalizing features.

Challenge 4 - Investigating the Impact of Contextual Knowledge for KG-DL: In the human brain, the same visual input is embedded differently in different brain areas depending on the context [5, 6, 7]. For instance, tools like a hammer or a screwdriver are embedded not only because of their appearance but also because of their function.

With KG-DL we are able to induce such additional context from a KG into a visual DL pipeline. Therefore, we expect that studying the role of context, i.e., the type of prior knowledge in KG, will also be crucial for further improving the performance of CV systems. The type of prior knowledge guides the DNN to focus on context-specific information in the dataset and learn appropriate features. A KG is a structured repository for any kind of prior knowledge, regularly contains heterogeneous types of prior knowledge that come from different sources and thus can be relevant in different compositions for different tasks.

It needs to be investigated, how the task-relevant contextual knowledge can be best extracted from the KG and how it can be best induced into a DNN. Further, the impact of context needs to be evaluated, especially for transfer learning tasks. Therefore, the relationship between the type of prior knowledge and the final performance in visual transfer learning needs to be analyzed, and relevant factors affecting it need to be identified and considered.

1.3 Research Questions

Based on the discussion in Section 1.1, we identified and elaborated problems and challenges for the topic of visual transfer learning using KG. In this section, we refine the research goal for the scope of this work and define four main research questions.

RQ1: Why can prior knowledge encoded in a KG improve DL-based visual transfer learning?

To understand how prior knowledge encoded in a KG can enhance DL-based visual transfer learning, we must delve into the fundamentals of transfer learning and explore various transfer learning scenarios. One key issue to investigate is why data-driven DL struggles with transfer learning and how transfer learning methods aim to address this issue. To do so, we must examine the specialization-generalization trade-off, the concept of inductive biases, and how they relate to prior knowledge more broadly. Moreover, we must define the notion of prior knowledge and its different forms. Additionally, we need to justify the use of KGs for encoding prior knowledge and outline the task of visual transfer learning using KGs in detail.

In summary, our research question requires a comprehensive analysis of the fundamentals of transfer learning, the concept of prior knowledge, and the use of KGs to encode and enhance transfer learning for visual data.

RQ2: What are possible ways of integrating the prior knowledge encoded in a KG into the DL pipeline?

To investigate possible ways of integrating prior knowledge encoded in a KG into the DL pipeline, we must consider the many attempts that have been made to combine these two modalities of knowledge. However, such integration is not always straightforward. Our research will summarize relevant approaches from various application domains and group them into distinct categories based on their underlying principles and infusion method. In creating these categories, we will focus on approaches that combine KGs with DL for computer vision as well as approaches that use other types of prior knowledge, such as text representations, in combination with visual data. Additionally, we

will divide each of these works into feature extractors and transformation models, based on their impact on the original encoder network.

To answer our research question in-depth, we must provide a detailed analysis of the approaches belonging to each category, highlighting their strengths and weaknesses, and discussing their potential applications to other KG-DL applications. Overall, our goal is to provide a comprehensive overview of the different ways to integrate prior knowledge encoded in a KG into the DL pipeline and to explore the potential benefits and limitations of each approach.

RQ3: How can prior knowledge encoded in a KG guide DL to improve visual transfer learning?

To understand how prior knowledge encoded in a KG can guide DL and improve visual transfer learning, a novel method that belongs to the *Knowledge Graph as a Trainer* must be developed. This will require a comprehensive study that describes the principles of the approach, the pipeline, and transfer learning experiments on multiple datasets. To develop the approach, a KG must first be created that contains prior knowledge about the domain and related domains. This involves exploring the type of graph structure used, the concepts used, and the best way to represent each class in the dataset. Additionally, a suitable infusion method for the knowledge of the KG into the learning process of the DNN must be found. Finally, the approach must be evaluated in transfer learning tasks such as domain generalization and domain adaptation against a baseline without a KG. The goal is to prove the hypothesis that the prior knowledge of a KG improves the generalization performance of DL methods.

In summary, to answer the research question, a new method will be developed that utilizes a *KG as a Trainer* for DNNs. The approach will be evaluated through transfer learning experiments, and the results will be compared to a baseline without a KG to determine the impact of prior knowledge on visual transfer learning.

RQ4: How does the type of prior knowledge in a KG impact KG-DL performance, especially in visual transfer learning?

Recent findings in cognitive science suggest that visual inputs are embedded differently in the brain depending on the context. To understand how the context, i.e. type of prior knowledge, in a KG impacts KG-DL performance, particularly in visual transfer learning, we need to investigate how specific contextual views of a generic KG influence a DNN. To train the DNN using a specific type of context, we need to extract a specific contextual view from the generic KG and combine it with the DNN in a way that the learned features from unstructured image data are influenced by the contextual view. We then need to investigate the results of the different learned contextual models

in depth, using qualitative visualizations of the class-based cosine similarities of their contextual embedding spaces. Additionally, we must analyze the approach quantitatively by comparing the final class accuracies on object recognition tasks for source and target domains.

To answer the research question, we need to investigate how the different contextual views of a KG influence the performance of a DNN for specific CV tasks. We must extract and combine the contextual views with the DNN and then evaluate the resulting performance through qualitative and quantitative analyses. By doing so, we can gain insights into how the type of prior knowledge in a KG can impact KG-DL performance in visual transfer learning.

1.4 Thesis Overview

To give the reader an overview of the work done in this thesis, we list the most important contributions of the thesis below. In addition, we provide all relevant publications and an overview of the structure of the work.

1.4.1 Contributions

This part summarizes the main contributions of the thesis on *visual transfer learning using knowledge graphs*. In the following, the contributions are listed based on the four main research questions.

1. *Structured analysis of why prior knowledge encoded in a KG can improve DL-based visual transfer learning.*

Contribution for RQ1. This contribution explores the suitability of KG for improving DL methods in visual transfer learning. The study provides an in-depth analysis of transfer learning fundamentals, explains the specialization-generalization trade-off of DL, and highlights how DNNs handle domain shifts using a raw format of prior knowledge. Additionally, the research defines transfer learning using KG, describes explicit and implicit prior knowledge, and argues why KGs are an ideal representation format for it.

2. *Categorization of methods combining KG and DL for visual transfer learning.*

Contribution for RQ2. This contribution presents a comprehensive categorization of KG and DL combinations for visual transfer learning. The study proposes four main categories of KG-DL combinations: *Knowledge Graph as a Reviewer*, *Knowledge Graph as a Trainee*, *Knowledge Graph as a Trainer*, *Knowledge Graph as a Peer*. This categorization framework provides a clear and structured understanding of the different ways in which KG and DL can be combined for visual transfer learning purposes.

3. *Method to learn a DNN using a KG as a Trainer.*

Contribution for RQ3. This contribution introduces a new method called KG-NN, which

uses a *KG as a Trainer* to train a DNN. The method works by transforming the explicit graph representation of the KG into vector space using a KGE method. KG-NN then uses a contrastive loss to train the DNN to align its implicit visual embedding with the domain-invariant embedding provided by the KGE. This method provides a novel approach to improve visual transfer learning using the knowledge encoded in KGs, ultimately enhancing the performance of DNNs in image recognition tasks.

4. *Method to learn contextual DNNs using contextual views of a generic KG.*

Contribution for RQ4. This contribution examines the impact of different types of prior knowledge encoded in a KG on the performance of KG-NN. The study analyzes the performance of KG-NN using individual contextual views, i.e. types of prior knowledge, of a generic KG and shows that the context in which the prior knowledge is presented influences the final accuracy of KG-NN, individual class accuracies, as well as predictions for individual image samples. This research demonstrates the importance of context in enhancing the accuracy of DNNs and highlights the potential benefits of using contextual views of KGs to improve visual transfer learning.

1.4.2 List of Publications

Part of the work in this thesis is based on the following publications:

1. **Sebastian Monka**, Lavdim Halilaj, Stefan Schmid, and Achim Rettinger. 2021. Learning Visual Models Using a Knowledge Graph as a Trainer. In *The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24-28, 2021, Proceedings (Lecture Notes in Computer Science)*, Springer, 357–373.
(*Chapters 3, 5*)
2. **Sebastian Monka**, Lavdim Halilaj, and Achim Rettinger. 2022. A survey on visual transfer learning using knowledge graphs. *Semantic Web* 13, 3 (2022), 477–510.
(*Chapters 2, 3, 4*)
3. **Sebastian Monka**, Lavdim Halilaj, and Achim Rettinger. 2022. Context-Driven Visual Object Recognition Based on Knowledge Graphs. In *The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings (Lecture Notes in Computer Science)*, Springer, 142–160.
(*Chapter 6*)
4. Juergen Luetttin, **Sebastian Monka**, Cory Henson, and Lavdim Halilaj. 2022. A Survey on Knowledge Graph-based Methods for Automated Driving. *KGSWC (2022)*.
(*Chapters 2, 3*)

5. Lavdim Halilaj, Juergen Luetttin, Cory Henson, and **Sebastian Monka**. 2022. Knowledge graphs for Automated Driving. In IEEE AIKE-Artificial Intelligence and Knowledge Engineering.
(Chapter 2)
6. Lavdim Halilaj, Juergen Luetttin, **Sebastian Monka** and Cory Henson. 2022. Knowledge graphs for Autonomous Driving. In IJCS 2023 - International Journal of Semantic Computing.
(Chapter 2)
7. Ruwan Wickramarachchi, Cory Henson, **Sebastian Monka**, Daria Stepanova, and Amit Sheth. 2022. Tutorial: Knowledge-infused Learning for Autonomous Driving (KL4AD). In Tutorial - ISWC 2022.
(Chapters 2, 3, 5, 6)

1.5 Outline of the Thesis

The thesis is comprised of seven chapters, each serving a specific purpose.

Chapter 1 provides a high-level introduction to the thesis by discussing the motivation for the topic of visual transfer learning using KG in Section 1.1. We formulate four main challenges in Section 1.2, define corresponding research questions in Section 1.3, and present the main contributions and relevant publications of this thesis in Section 1.4.

Chapter 2 provides the necessary background information for KG-DL and NSAI in Section 2.1, related to the deeper-rooted question of why to combine symbolic AI and machine learning. This chapter also introduces the three thesis-relevant modalities of data, information, and knowledge, namely vision, language, and KG in Section 2.2. It discusses the feature extraction process in Section 2.3, and inductive biases in DL in Section 2.4.

Chapter 3 defines the term transfer learning using prior knowledge and answers **RQ1**. Therefore, it introduces the fundamentals of transfer learning in Section 3.1. The chapter continues in Section 3.2 with insights into the problem of DL with domain change, introduces the specialization-generalization trade-off, and describes how recent DL methods try to deal with transfer learning tasks. Further, Section 3.3 introduces transfer learning using KG, by presenting KGs as an ideal representation format for prior knowledge. In addition, Section 3.4 provides resources for visual transfer learning using KG, including KGs and image datasets with prior knowledge.

Chapter 4 contains a structured analysis of how KGs can be used with DL methods. Therefore, it provides a summary and categorization of methods for KG-DL combinations for visual transfer learning and answer **RQ2** in Section 4.2 by introducing four main categories of how a KG can be combined with DL. Further, Section 4.3 summarizes relevant surveys related to the chapter.

Chapter 5 answers **RQ3**, by proposing a method for learning a DNN using a *KG as a Trainer* in Section 5.1. We show in Section 5.2 that the proposed method outperforms baselines without a KG,

especially on visual transfer learning tasks. Section 5.3 provides an overview of related work.

Chapter 6 answers **RQ4** by presenting a method for learning contextual DNNs using contextual views of a generic KG in Section 6.2. Accordingly, this chapter investigates the effect of the type of prior knowledge in the KG on the final results of transfer learning tasks for object recognition in Section 6.3. Section 6.4 summarizes related work relevant to the chapter.

Finally, Chapter 7 concludes the thesis with a discussion of the results, challenges, and potential applications for future work.

Chapter 2

Preliminaries

This chapter gives an overview of relevant prior information on visual transfer learning using KG. Therefore, NSAI is introduced based on theories of human intelligence, the modalities of vision, language, and KG relevant for the work are introduced, the DL-based feature extraction process and the idea of embedding spaces are explained, and insights into the inductive biases of DL that influence what a DNN learns are provided.

Most of the topics described in this chapter are already published in:

- **Sebastian Monka**, Lavdim Halilaj, and Achim Rettinger. 2022. A survey on visual transfer learning using knowledge graphs. *Semantic Web* 13, 3 (2022), 477–510.

The chapter begins by deriving the idea of NSAI from theories of human intelligence in Section 2.1. For this purpose, we analyze views from cognitive science, e.g., philosophy, psychology, and neuroscience, link them to the distinct principles of symbolic AI and ML, and introduce the combination of KG and DL as a representative of NSAI. In the context of this work, we define vision, language, and KG as modalities of data, information, and knowledge in Section 2.2. Finally, we explain the principle of feature extractors in Section 2.3, their embedding spaces, and the general effect of inductive biases in DL in Section 2.4.

2.1 Neuro-Symbolic AI

Enabling machines to behave intelligently like humans is one of the main challenges of AI. Human intelligence is often defined as the ability to achieve goals in a wide range of environments [8]. Whereas ML approaches achieve impressive results on several tasks, they can only solve problems in their specific training domains. With the goal of improving ML, other approaches to AI, like symbolic AI, came back into focus. Symbolic AI and ML have evolved separately over the years.

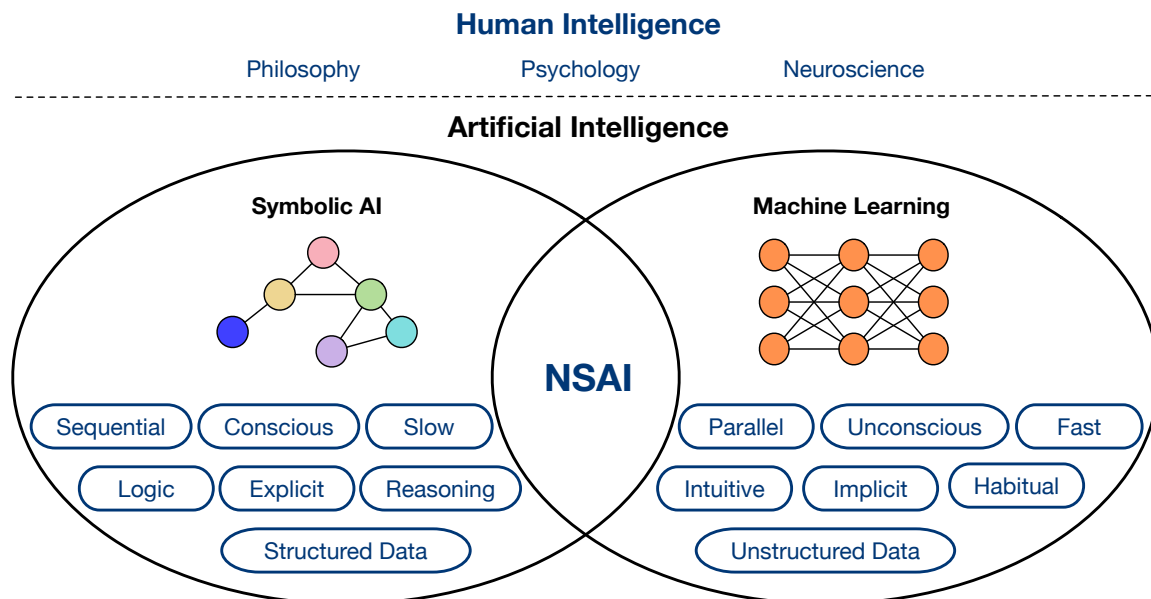


Figure 2.1: Prerequisites for artificial intelligence can be derived from theories of human intelligence, based on research in philosophy, psychology, and neuroscience. Human intelligence is comprised of two working processes: a sequential, conscious, and slow part, based on logic and reasoning, forming explicit representations of structured data, akin to the principles of symbolic AI. The parallel, unconscious, and fast part, based on habitual and intuitive processing, forms implicit representations of unstructured data, akin to the principles of machine learning.

A KG is part of the symbolic AI, hence they make use of structured data, logic, and reasoning, and operate in an explicit and sequential manner. DL, however, is part of the ML approaches. It can leverage knowledge from unstructured data, is fast, intuitive, and operates in an implicit and parallel manner. Because of their complementary properties, recent efforts have been made to combine the two fields into one called NSAI. In this new form of AI, ML is used to develop a deep understanding of unstructured data, and symbolic AI is used to provide structure, causality, and interpretability.

Figure 2.1 illustrates the relation between symbolic AI and ML and their relationship to human intelligence and AI. Nowadays, NSAI mainly uses symbolic AI and ML as sequential and independent modules. However, cognitive science presumes that both fields should be connected in a hybrid and interactive manner [9]. Therefore, NSAI needs to draw inspiration from cognitive theories of human decision making [10], by enriching DL approaches with well-founded knowledge representations and reasoning capabilities [11].

In this chapter, we shed light on why a combination of both fields seems promising. Therefore, we introduce theories of human intelligence from philosophy, psychology, and neuroscience and link them to the properties of symbolic AI and ML. Further, we introduce the combination of KG and DL as a promising step to build NSAI models that suit the complexity of the real world.

2.1.1 Human Intelligence

Defining human intelligence is a widely discussed and controversial topic [12]. There are efforts from philosophy, psychology, and neuroscience to better understand the fundamentals of intelligence. These disciplines differ in their focus and approach to understanding the human mind and behavior. Whereas philosophy uses a more abstract and theoretical approach, psychology and neuroscience are empirical and data-driven. Psychology focuses on the study of behavior and mental processes, while neuroscience focuses on the biological origin of these processes.

Philosophy: Philosophers in ancient Greece already started to think about human intelligence. Due to Plato, intelligence is the perceptible part of the soul, and knowledge is justified true belief. He separated intelligence into discursive reason and intuitive reason [13]. The discursive reason is based on explicit steps, it is slow and based on logic and rules. Intuitive reason is a fast use of reason, drawing conclusions without going through the whole deductive process [14]. Aristotle distinguished between three types of souls, where the intellectual soul corresponds to humans and their ability to think [15]. Later, Descartes, Félix Ravaisson, and Henri Bergson defined intuition.

In general, philosophers described the brain as a system that uses symbolic logic to reason about the world [16]. If the mind goes beyond the data given, another source of information must make up the difference [17].

Psychology: In the 20th century, intelligence became an object of study in psychology. Jean Piaget studies the development of intelligence in children. Howard Gardner proposes a popular but controversial theory of multiple intelligences. Catell introduced 1971 the theory of fluid and crystallized intelligence [18, 19].

Aligned with assumptions from philosophy [13], Kahneman [9] introduced a framework about human intelligence. He distinguished between two working principles, a *system 1* type and a *system 2* type. *System 1* is relatively slow, uses logic and reasoning, and operates in an explicit and sequential manner. *System 2* is fast, intuitive, and operates in an implicit and parallel manner. Moreover, *system 1* and *system 2* work in tandem, not as separate entities.

In psychology, it is assumed that the large majority of the skills and knowledge is learned, rather than innate [20]. However, some kind of core knowledge seems to be built into the structure of the brain, since human knowledge comes with priors [8].

Neuroscience: Neuroscience, as a descendant of philosophy and psychology, became an independent research field in the 20th century. In 1952 the first version of a biological *spiking neural network* (SNN), the Hodgkin–Huxley model, was introduced to explain the working principles of a network of neurons in the brain [21]. Whereas an SNN differs from the traditional artificial NN from computer science in many ways, their abstract working principles are comparable. Referring

to the findings of psychology [9], neuroscientists claim that the brain is not literally divided into two systems. Moreover, these systems evolve based on higher-level abstractions. *System 1* and *system 2* work in tandem, not as separate entities, such that a single reasoning process always involves both systems. It is shown that there is no better system, both systems can be biased and make mistakes.

However, we do not yet know enough about neuroscience to literally reverse engineer the brain. AI can help to decipher the brain, rather than the other way around [22]. Since most findings are based on models comparing the brain with artificial NNs, these findings need to be taken with caution.

2.1.2 Artificial Intelligence

One of the ultimate goals of AI is to develop systems that possess intelligence similar to that of humans. Therefore, most AI methods draw inspiration from theories of human intelligence. While symbolic AI is oriented to the brain's sophisticated reasoning capabilities that use symbolic processing, ML focuses on the unconscious and intuitive part, where information is extracted from a variety of unstructured data. In the history of AI, there have been many debates about which AI approach is the most promising.

Symbolic AI: Symbolic AI focuses on human reasoning capabilities. Symbolic logic can be traced back to the rhetoric of ancient Greece, which was later developed into formal logic. The goal of formal logic was to provide a complete and rigorous foundation for mathematics, but also to produce logical representations of everyday knowledge. Nowadays, instead of the strings or trees of classical logic, graphs are increasingly used to represent knowledge. These KGs are vast graph networks that are increasingly finding applications in web search and product recommendation [16].

In general, symbolic AI is based on logic and declarative knowledge representation. It mimics human reasoning, which is why the methods are explicit and understandable to humans. Moreover, these methods thus provide an interface to the representation of expert knowledge. The approaches work with symbols, which are high-level abstractions of real-world information and are therefore able to generalize to unknown domains.

However, reasoning on symbols also has some drawbacks. For example, symbolic AI approaches are quite slow because they are sequential in nature. They require symbols based on structured data as input, which are very difficult to determine in unstructured data such as images or texts. Symbolic AI methods are therefore mostly used in controlled environments and include logic, planning, search, and optimization.

Machine Learning: ML is based on the idea of developing systems that, like humans, learn from observations and can thus draw conclusions about their environment. In the course of research into the learning behavior of living things, biological learning theories were formulated as early as around

1940 [23, 24]. These mathematical principles were later first implemented in the perceptron [25] algorithm, in which a single neuron is trained. With the invention of the backpropagation [26] algorithm, it was possible to train deeper neural networks with one or two hidden layers, which were able to solve more complex tasks because of their higher capacity. With the further increase of neurons in additional hidden layers, data, and computational power, DL approaches [27, 28, 29] have been able to provide significant results for a range of tasks. Moreover, the wide application of DL in various areas of AI has earned ML a high reputation in society.

ML approaches can work with unstructured data, such as text and images. Based on raw data, they can extract recurring patterns and classify or predict them. Their strength is that they learn a model that works well in the complexity of the real world. However, they lack interpretability and explainability. They require large and well-curated datasets and have limited generalizability and robustness.

2.1.3 Combining Knowledge Graph and Deep Learning

Based on theories of human intelligence, as well as the complementary strengths and weaknesses of symbolic AI and ML, there is a recent trend toward NSAI. In particular, the combination of KGs, as a method of symbolic AI, and DL, as a method of ML, led to promising results.

First, the symbolic principle of the human mind, a sequential, slow, and deliberate processing of explicit, structured knowledge involving logic and reasoning, can be related to the properties of a KG. KGs work with structured data and symbolic representations, use sequential processing, and are relatively slow. They use rules, reasoning, and compositionality based on and driven by general human knowledge, resulting in general knowledge modules that can be reused for related domains.

Second, the subsymbolic principle is a parallel, rapid, and unconscious processing of implicit, unstructured knowledge that is intuitive and habitual. This principle can be associated with DL, which extracts information from unstructured data, such as text or images, and is highly parallel and fast. However, they are implicit and therefore not understandable and interactable by humans.

Although all fields of research agree that human intelligent behavior is reflected in a synergy of symbolic and subsymbolic behavior, it is still not clear whether the symbolic structure is innate or evolves from the implicit knowledge of a densely interconnected biological NN. This debate can be directly applied to the combination of KG and DL. While some argue that symbolic behavior should be explicitly modeled [30], others think that symbolic structures evolve from the DL itself [31]. These discussions raise further questions, e.g. how a future NSAI architecture should look such that the explicit knowledge in the KG can be used efficiently with DL.

2.2 Modalities of Data, Information, and Knowledge

The goal of AI is to enable machines to interpret and understand the world in the same way that humans do. This includes the ability to automatically analyze raw data, convert it into information about objects, and make predictions about the world.

The terms data, information, and knowledge are closely related but also have significant differences. A widely accepted definition for data, information, and knowledge is that data are character sets representing empirical stimuli or perceptions, information is a character set representing empirical knowledge, and knowledge is a character set representing the meaning of thoughts that the individual reasonably believes to be true [32]. In short, data is the raw material, information is the processed material, and knowledge is the understanding that results from processing the information.

Implicit knowledge is the form of knowledge that is gathered through experience and repetition and is hard to articulate or transfer to others. Explicit knowledge is the type of knowledge that is conscious. It can be formalized, interpreted, managed, and transmitted to others.

Moreover, data, information, and knowledge can occur in different modalities, i.e. specific forms of representation. For instance, humans work with modalities such as vision, audition, touch, etc. In the scope of this thesis, we focus on a more fine-grained concept of modality. Hereby, the modality of vision relates to visual representations, the modality of language to language representations, and the modality of KGs to KG representations. For the task of visual transfer learning, we further define the sources of prior knowledge, i.e. language and knowledge graph, as semantic modalities.

2.2.1 Vision

The modality of vision refers to data, information, and knowledge that is based on images. The generation of images or digital images is based on the fundamentals of the human visual system. The human visual system processes the light information in the eye and encodes the visual signal into electrical stimuli using neurons on the retina. Likewise, camera sensors record the visual information of the world in digital images. Most sensors transform the visual signal of the world into a continuous voltage waveform. This waveform is then converted into a digital image by sampling and quantization. Therefore, an image is not only defined by the visual input but also by the sensor and its digitization procedure.

A digital image is composed of picture elements, also called pixels. Each pixel has a finite, discrete set of numerical representations for its intensity or gray level, which is an output of its two-dimensional functions input by its spatial coordinates denoted by x , y , respectively [33]. Since images exist in a fixed-size grid structure in Euclidean space, images can be seen as a particular case of graphs with special adjacency [34].

In addition to grayscale information, images can also encode color information. Color images can be represented using various color models, e.g., RGB, CMYK, or HSI. In Figure 2.2 we show

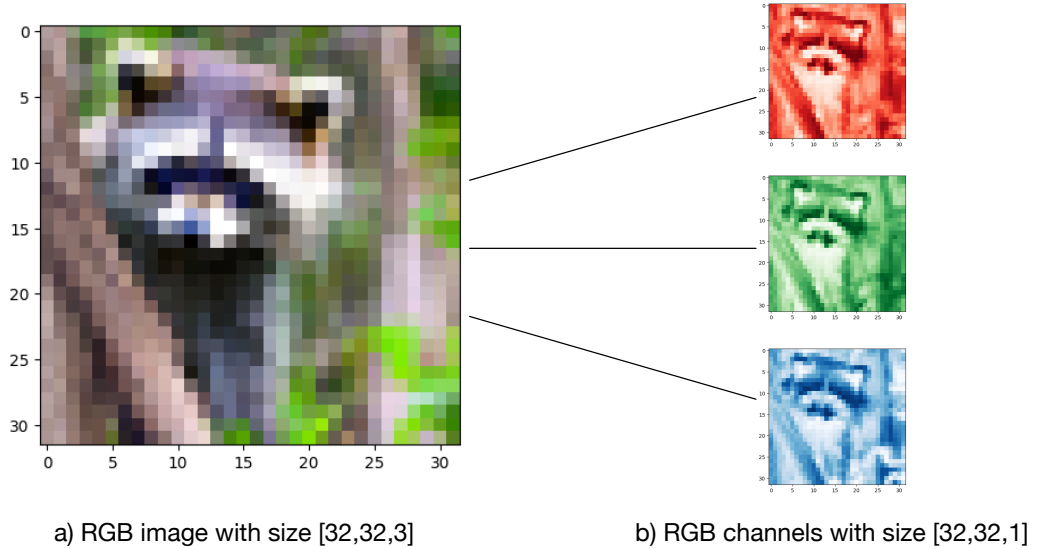


Figure 2.2: The modality of vision is an encoded version of the visual signal based on the light information in the real world. An RGB image is a digitized version of the visual signal and consists of pixels, here 32x32, and three RGB channels, i.e. red, green, and blue. The information in the image is comprised of the individual pixels of the channels and their correlations.

the discrete nature of an RGB image. An RGB image consists of discrete pixels with three channels each for the colors red, green, and blue.

Besides the structure of an image, it is also relevant what kind of information is contained in an image. The information of an image is defined by regions or parts of the image. These regions are called features and describe, among other things, color, shape, and texture [35]. Based on these features, humans, but also machines, are able to derive higher-level semantic information, e.g. objects or classes, from an image.

2.2.2 Language

The modality of language includes data, information, and knowledge that is based on texts. In the scope of visual transfer learning, we refer to the modality of language as a semantic modality. Humans use language data as a medium for communication and knowledge transfer.

Linguistics is the scientific study of human language and text, that investigates the structure of sentences (syntax), their meaning (semantics), the structure of words (morphology), speech sounds and equivalent gestures in sign languages (phonetics), the abstract sound system of a particular language (phonology), and how social context contributes to meaning (pragmatics) [36]. Text linguistics is a subfield of linguistics that studies language used in written or spoken texts [37]. It provides linguistic theories, concepts, and frameworks that are used to analyze text data. The subfield of computational linguistics deals with the use of computer technology to analyze and generate natural

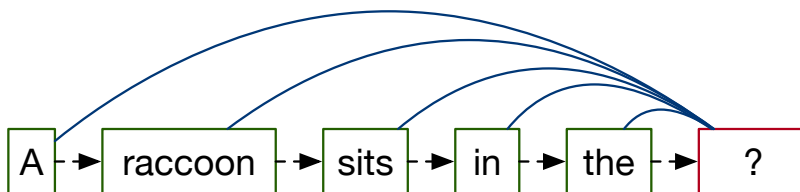


Figure 2.3: The modality of language has a rich underlying structure based on syntax, semantics, morphology, etc. A sentence is an extended structure of syntactic units. It consists of many features interlinking the statistical probability of a word (?) occurrence to the previous words and their order.

language. It provides the technical tools and methods to apply linguistic theories and concepts to large-scale text data.

As illustrated in Figure 2.3, text data is an extended structure of syntactic units such as words, groups, and clauses and textual units that is marked by both coherence among the elements and completion. A non-text consists of random sequences of linguistic units such as sentences, paragraphs, or sections in any temporal and spatial extension. [38] Practically, text data is a sequence of discrete tokens [39] and can be seen as a one-dimensional grid or ring graph [34]. Language is therefore a finite collection with its complexity limited by the number of letters, tokens, and words that can be represented.

To extract information from text data, rules or statistical methods are applied to extract text features. These features are, for instance, the grammatical structure, the used characters and their frequency, and co-occurrences. Based on these features, humans, as well as machines, can derive higher-level sets of information, such as meaning.

2.2.3 Knowledge Graph

The modality of KG describes data, information, and knowledge that is based on KGs. In the scope of visual transfer learning, we refer to the modality of KG as a semantic modality. A KG is a form of a semantic network, that stores explicit knowledge in a graphical format [40]. KGs are a structured representation of facts, consisting of entities, relationships, and semantic descriptions. A comprehensive definition is given by the authors of [41] where a KG is defined as *a graph of data with the objective of accumulating and conveying real-world knowledge, where entities are represented by nodes and relationships between entities are represented by edges*. Entities can be real-world objects and abstract concepts, relationships represent the relation between entities, and semantic descriptions of entities and their relationships contain types and properties with well-defined meanings.

Knowledge can be expressed in a factual triple in the form of (head, relation, tail) or (subject, predicate, object) under the Resource Description Framework (RDF), for example, (Albert Einstein, WinnerOf, Nobel Prize). In its most basic form, we see a KG as a set of triples $G = H, R, T$, where

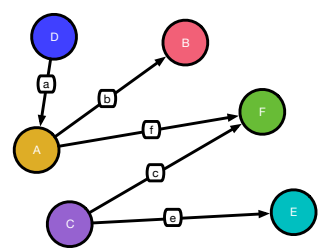
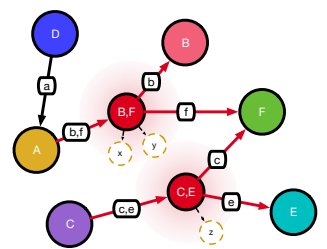
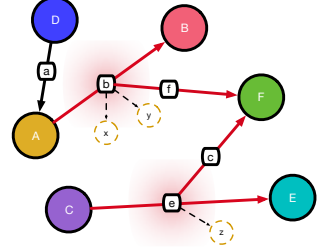
	Directed Labeled Graphs	Hypergraphs	Hyper-Relational Graphs
Nodes and Literals	- Real-world and abstract entities - Entity's attribute value	- Real-world and abstract entities - Entity's attribute value	- Real-world and abstract entities - Entity's attribute value
Relationships	- Binary relations between entities - Relations between an entity and its attribute's values	- Binary relations between entities - Relations between an entity and its attribute's values - Many-to-many relations between entities (Hyperedge)	- Binary relations between entities - Relations between an entity and its attribute's values - Additional information encoded in a relationship (Hyper-relation)
Semantics	Connect two nodes	Connect an arbitrary set of nodes	Connect two nodes with additional contextual information
Example			

Table 2.1: A KG can be constructed using various graph models. The following three common graph models can be used as an underlying structure for knowledge representation in KGs: 1) Directed Labeled Graphs; 2) Hypergraphs; and 3) Hyper-relational Graphs.

H is a set of entities, $T \subseteq E \times L$, is a set of entities and literal values and R , set of relationships which connects H and R .

A graph model is a model which structures the data, including its schema and/or instances in form of graphs, and the data manipulation is realized by graph-based operations and adequate integrity constraints [42]. Each graph model has its formal definition based on the mathematical foundation, which can vary according to different characteristics, for instance, directed vs. undirected, labeled vs. unlabeled, etc. The most basic model is composed of labeled nodes and edges, easy to comprehend but inappropriate to encapsulate multidimensional information. Other graph models allow for the representation of information utilizing complex relationships in the form of hypernodes or hyperedges.

Table 2.1 illustrates three graph models, namely directed labeled graphs, hypergraphs, or hyper-relational graphs, a KG can be constructed. A KG can be based on any such graph model utilizing nodes and edges as a fundamental modeling form. To extract information or knowledge from the raw graph data, methods operate on symbolic rules or ML. Both approaches extract features in the graph data, based on regions or pieces. Such features can be for example specific combinations of nodes and relations, frequencies of their occurrences, or distinct graph structures. In the following, we discuss in detail the three common graph models used in practice to represent data graphs:

Directed Labeled Graphs: A directed labeled graph is comprised of a set of nodes and a set of edges connecting those nodes, labeled based on a specific vocabulary [42]. The direction of the

edge of two paired nodes is important, which clearly distinguishes between the start node and the end node. This intuitively enables the organization of information via the utilization of binary relationships.

Hypergraphs: Hypergraphs extend the definition of binary edges by allowing the modeling of multiple and complex relationships [42]. On the other hand, hypernodes modularize the notion of a node, by allowing nesting graphs inside nodes. In addition, the notion of a hyperedge enables the definition of n-ary relations between different concepts.

Hyper-Relational Graphs: A hyper-relational graph is also a labeled directed multigraph where each node and edge might have several associated key-value pairs [43]. Internally, nodes and edges are annotated according to a chosen vocabulary and have unique identifiers, making them a flexible and powerful form of modeling for graph analysis with weighted edges.

2.3 Feature Extraction

A feature extractor is a transformation function from higher dimensional into lower dimensional vector space [44, 45]. Since it has been shown that most downstream tasks can be solved better on a reduced dimensionality, feature extractors are also a fundamental building block of modern systems working on visual and semantic data.

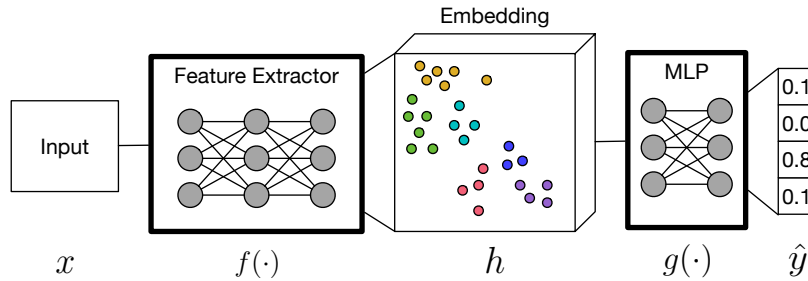


Figure 2.4: A DNN that takes \mathbf{x} as input and predicts $\hat{\mathbf{y}}$ can be decoupled into a feature extractor $f(\cdot)$ with its embedding space \mathbf{h} and a projection head $g(\cdot)$.

However, more and more conventional feature extraction methods have been replaced with DNNs. A DNN is an artificial *neural network* (NN) with multiple layers between the input and output layers, having the ability to automatically extract lower dimensional features from the input data [46, 47]. As depicted in Figure 2.4, a DNN can be decoupled in a feature extractor $f(\cdot)$, with its embedding \mathbf{h} and a projection head $g(\cdot)$, expressing the function

$$\hat{\mathbf{y}} = g(f(\mathbf{x})). \quad (2.1)$$

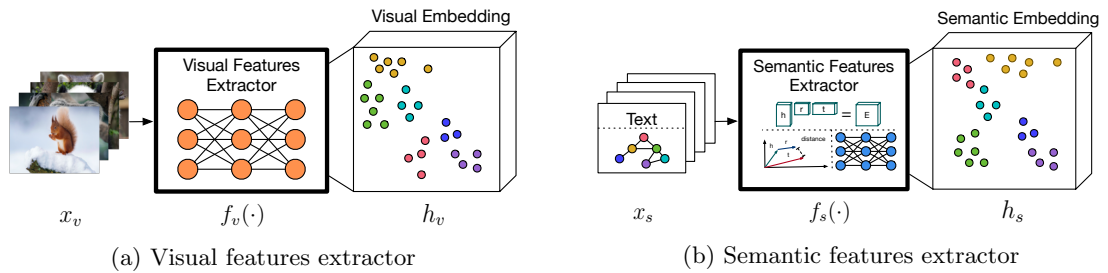


Figure 2.5: A Feature extractor transforms input data into embedding space: a) a visual features extractor transforms visual input data, i.e. images, into visual embedding space; and b) a semantic features extractor transforms semantic input data, e.g. text or graphs, into semantic embedding space.

There are different architectures of DNNs, but they always consist of the same components: neurons, synapses, weights, biases, and functions [48].

The basic architecture that builds a DNN is a *multilayer perceptron* (MLP). An MLP can be extended by modules using convolutions, recurrences, or attention. Whereas DNNs are usually trained end-to-end resulting in a task-dependent embedding space \mathbf{h} , more recently, attempts have been made to independently pre-train the feature extractor and their embedding space, such that it can be applied to several visual transfer learning downstream tasks [49].

Embedding: When referring to an embedding or embedding space we address a vector-based representation of input data. A feature extractor maps input data \mathbf{x} to an embedding $\mathbf{h} = f(\mathbf{x}) \in \mathbb{R}^{d_E}$, where the activations of the final pooling layer and thus the representation layer have a dimensionality d_E , where d_E depends on the architecture of the feature extractor that is used. Ideally, an embedding captures the semantics of the input by placing semantically similar inputs close together in the vector space.

Projection: A projection or projection space is a transformation from the embedding to a different dimensional vector space. This transformation is done in DNNs using a projection head $g(\mathbf{h})$, e.g. MLP. A *projection head* $g(\cdot)$ maps the embedding \mathbf{h} into a projection $\mathbf{z} = g(\mathbf{h}) \in \mathbb{R}^{d_P}$. For the projection network $g(\cdot)$, often a multi-layer perceptron [50] with an input dimensionality d_E , and output vector of size d_P is used. In standard end-to-end DNN training for classification, d_P is the number of predicted classes and $\mathbf{z} = \hat{\mathbf{y}}$.

2.3.1 Visual Features Extractor

As shown in Figure 2.5, feature extractors can extract different types of information and therefore occur in different forms. A visual features extractor $f_v(\cdot)$, shown in Figure 2.5a, is a transformation function that transforms visual input data \mathbf{x}_v from a higher-dimensional image space into a lower

dimensional visual embedding space \mathbf{h}_v . A formal definition is given by

$$\mathbf{h}_v = f_v(\mathbf{x}_v), \quad (2.2)$$

where the final dimensionality of \mathbf{h}_v is determined by the architecture.

Whereas early approaches used traditional visual features extractors as *scale-invariant feature transform* (SIFT)[51] or *histogram of oriented gradients* (HOG) [52], modern CV methods use almost only DNN-based approaches. Architectures of DNNs to process images are MLPs, *convolutional neural networks* (CNN), *recurrent neural networks* (RNN), and *transformers*.

Each architecture has its advantages and is therefore preferred for a particular type of input data and particular task [48]. MLPs use fully connected neurons between each layer of the network, CNNs move filter kernels over grid-structured input data to learn recurrent features and thus scale to large and computationally intensive images, RNNs use recurrent connections to process temporal information of the input data [48], and transformer models use a multi-headed attention mechanism to learn rich information of the underlying input data [53].

Visual Embedding: A visual embedding contains low-dimensional representations of images. Each image is therefore associated with a particular vector in the visual embedding. It has been shown that the embedding encapsulates relevant information about the images in a dense format. Therefore, the similarity of images is encoded by distances in the embedding. In a visual embedding learned from data, visually similar images are more likely to be encoded near, while visually dissimilar images are more likely to be encoded far. It is quite common to use pre-trained visual feature extractors for large image datasets and use them for related tasks.

2.3.2 Semantic Features Extractor

A semantic features extractor $f_s(\cdot)$, shown in Figure 2.5b, is a transformation function that transforms semantic input data \mathbf{x}_s from a higher dimensional image space into a lower dimensional semantic embedding space \mathbf{h}_s . A formal definition is given by

$$\mathbf{h}_s = f_s(\mathbf{x}_s), \quad (2.3)$$

where the final dimensionality of \mathbf{h}_s is determined by the architecture.

The term semantic data is here used for both, unstructured data from language and structured data from a KG. Although the input data structure differs in its original format, the output of the semantic features extractor is always a low-dimensional and vector-based semantic embedding space. This similarity enables a seamless transfer from hybrid approaches of vision and language to hybrid approaches of vision and KGs.

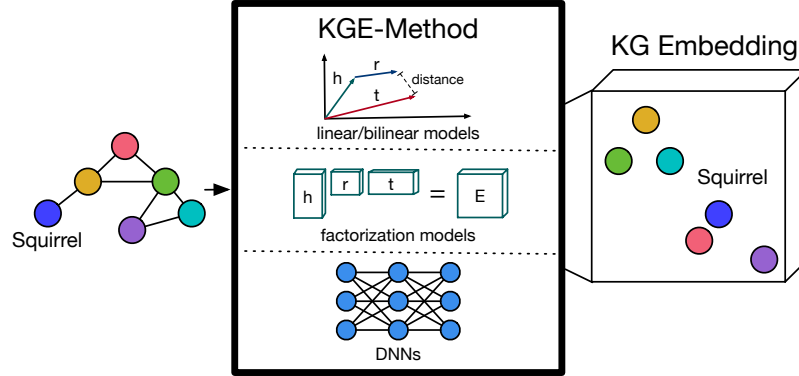


Figure 2.6: A KGE method transforms a *knowledge graph* (KG) into a vector-based KG embedding \mathbf{h}_s such that each node is assigned to a position in vector space. There are many different types of KGE methods, e.g. linear/bilinear models, factorization models, or DNNs.

Knowledge Graph Embedding: A knowledge graph embedding \mathbf{h}_s is a representation of a KG in vector space, where close relationships between entities in a KG are reflected by local neighborhoods. KGE is generated by a KGE method $KGE(\cdot)$, which maps the entities and relations of a KG into low-dimensional vectors while capturing their semantic meanings and relations [54]. Therefore, a KGE method is a special case of the semantic features extractors $f_s(\cdot)$ that works on graph data.

In Figure 2.6, the general pipeline of KGE methods which transform a KG into \mathbf{h}_s is illustrated. KGE methods can be categorized based on their learning mode and their used input format. Since most graph embedding algorithms are originally developed on tripled-based entity relations before being extended to n-ary relational data, we further divide the KGE methods into KGE methods that work on directed labeled graphs, hyper-relational graphs, and hypergraphs as introduced in Section 2.2.3.

KGE Methods - Learning Mode: Originally, KGE methods were developed to solve graph-based tasks such as node classification or link prediction. However, there is an increasing interest to apply KGE methods for visual tasks, such as classification, detection, or segmentation. We briefly categorize KGE methods therefore into unsupervised and supervised KGE methods, as Chami et al. [55] recently proposed for graph embedding algorithms.

Unsupervised KGE methods form \mathbf{h}_s based on the inherent graph structure and the node features, without considering additional task-specific labels for the graph or its nodes. An overview of unsupervised KGE methods is given by Ji et al. [56], who categorized KGE methods based on their *representation space* (vector, matrix, and tensor space), the *scoring function* (distance-based, similarity-based), the *encoding model* (linear/bilinear models, factorization models, neural networks), and the *prior information* (text descriptions, type constraints).

In contrast, supervised KGE methods learn \mathbf{h}_s to best predict node or graph labels. Forming

\mathbf{h}_s by using task-specific labels for the node features, \mathbf{h}_s can be optimized for a particular task while retaining the full expressiveness of the graph. The most common supervised KGE methods are *graph neural networks* (GNN) [57]. GNNs are extensions of standard DNNs that can directly work on a graph structure as provided by a KG. For scalability reasons and to overcome challenges that arise from graph irregularities various adaptations have emerged, such as *graph convolutional networks* (GCN) [58] or *graph attention networks* (GAT) [59]. Furthermore, non-Euclidean graph convolutional methods, such as *hyperbolic graph convolutional neural networks* (HGNCN) [60] are used to deal with a hierarchical structure of the input data.

KGE Methods - Input Type: The majority of existing KGE methods only work on directed labeled graphs, expecting binary relations in a tripled-based format [61, 62, 63]. However, as shown in Section 2.2.3, a basic triplet representation oversimplifies the complex nature of the information that can be stored in hypergraphs and hyper-relational graphs [64].

A hypergraph or hyper-relational graph can be transformed into directed labeled graphs, either by *reification* [65], that converts the graphs into binary-relation graphs, by creating additional triplets from a hyper-relational fact or by the *star-to-clique* [66] technique, that converts a tuple defined on k entities into $\binom{k}{2}$ tuples. However, these conversions lead to suboptimal and incomplete models as well as information loss. They only convert a set of key-value pairs, that are unaware of the triplet structure [64, 65]. To preserve the whole expressiveness of the KG, a set of new KGE methods are developed to directly operate on hypergraphs and hyper-relational graphs.

Methods that are extended to deal with hypergraphs are HEBE [67], HGE [68], Hyper2vec [69], HNN [70], HCN [71], DHNE [72], HHNE [73], Hyper-SAGNN [74], HypE [65]. HEBE [67] aims to learn the embedding for each object in a specific heterogeneous event by representing it as a hyperedge and HGE [68] incorporates multi-way relations using Laplacian tensors. Hyper2vec [69], hypergraph neural networks [70] and hypergraph convolution networks [71] are used for hypergraph embedding, however, cannot be directly used for predicting hyperedges. In addition, DHNE [72] and HHNE [73] use an MLP as their encoding model and therefore are limited to a fixed type and size of heterogeneous hyperedges. Hyper-SAGNN [74] is based on a self-attention-based graph neural network that can predict hyperedges and deal with variable hyperedge size and HypE [65] learns positional convolutional embeddings for each entity.

Methods that embed hyper-relational graphs are m-TransH [66], HypE, HSimple [65], RAE [75], GETD [76], TuckER [77], NaLP [78], HINGE [64], StarE [79]. m-TransH [66] that uses the star-to-clique method to convert the graph to binary relations. HypE, HSimple [65] and RAE [75] convert the hyper-relational facts into n -ary facts with one abstract relation that represents all relations of the original fact. GETD [76] and TuckER [77] are tensor factorization approaches for n -ary relational facts. NaLP [78] and HINGE [64] are convolutional models that support multiple entities and relations in one fact. Recently, GNNs have been demonstrated to be capable of encoding also multi-relational KGs [80], such as StarE [79] that can encode the structure of hyper-relational graphs.

2.4 Inductive Biases in Deep Learning

When data and computation increase a definition of heuristic specifications becomes harder and training learning systems become more effective [81]. However, all data-driven methods such as DL use inductive biases or prior assumptions to learn a model from data, as depicted in Figure 2.7. For instance, they assume the smoothness of the hidden function, the translation invariance for CNNs, and a hierarchical organization principle of concepts. The no-free-lunch theorem for ML [82] basically says that some set of priors over the space of all functions is necessary to obtain generalization.

However, priors can be imperfect and therefore they needed to be collected with care [83]. KGs and other structures such as probabilistic variants of those are only one way to inject human knowledge into a model [4]. Therefore, we list inductive biases that influence intentionally and unintentionally the learning of the model. Previous work [84] describe levels of knowledge integration, based on the training data, the hypothesis set, the learning algorithm, and the final hypothesis. We explain how the labels, the data augmentation, the network architecture, and the loss function influence the learned DNN.

2.4.1 Dataset

The dataset on which the DNN is trained can be seen as an inductive bias. It provides the model with a set of examples to learn from, and these examples shape the model’s internal representation of the problem domain. The specific patterns and regularities that are present in the dataset influence how the DNN learns from data and therefore how it generalizes to new data.

DNNs are data-driven approaches, which means that they adapt highly to the image data distribution of the training dataset. Therefore, the underlying independent and identically distributed (i.i.d.) assumptions of machine learning require the training domain to perfectly reflect the test domain. However, this strong assumption does not hold for real-world scenarios, where the test domain always differs from the training domain. Features that do occur in the training data but are not present in future testing domains are called spurious correlations.

DNNs are known to learn many spurious correlations of a training dataset, such as camera metadata or background information affecting the specific task [85]. For example, if a dataset contains a lot of images of dogs in a forest, a DNN that is trained on this dataset may learn to associate dogs with forests, rather than learning the higher-level semantics of what a dog is.

Learning spurious correlations of the dataset limits the applicability of DNNs because they do not reflect the higher-level semantics of the images. To avoid learning spurious correlations, it is important to have a balanced dataset. We further argue that a dataset needs to have a balanced distribution of samples and a balanced distribution of style to learn generalizable features that can be applied in reality.

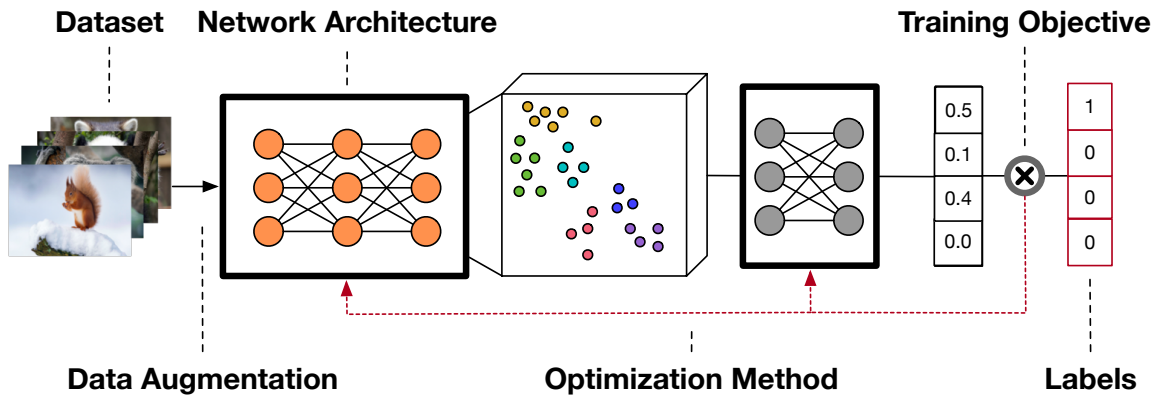


Figure 2.7: Inductive biases in deep learning. How a *deep neural network* (DNN) learns from data heavily depends on several inductive biases, such as the dataset, the data augmentation, the labels, the network architecture, the training objective, and the optimization method. Therefore, the combination of these inductive biases determines the performance of a DNN.

Distribution of Samples: Statistical learning methods, e.g. DL approaches, highly depend on the number of training samples for individual classes. Having a balanced dataset ensures prediction fairness. If the dataset is not balanced, low frequent classes will not be learned since their occurrence probability is too low. However, often low-frequent classes are important to be detected. For instance in autonomous driving, corner case scenarios contain safety-relevant information, such as dangerous traffic situations. Therefore, corner case scenarios are artificially collected or re-weighted to ensure their detection even if the data is underrepresented.

Distribution of Style: The style of available image datasets varies greatly. Most image datasets are collected from a variety of publicly available sources on the Internet [86]. Therefore, they contain many different styles of images, e.g. different color distributions, lighting, resolution, etc. This enables large image datasets to regularize their models in some sense. However, in autonomous driving, most datasets are collected through driving scenarios with a similar camera setup. For example, a DNN that was trained on a fisheye camera will fail if applied to a linear camera, or a model trained on grayscale images will lose performance if applied to color images.

2.4.2 Data Augmentation

Data augmentation is the process of changing the distribution of the input data. It is a method to artificially increase the number of data samples, but also a way to regularize a model by inducing some prior knowledge about the test domain. The goal is to avoid overfitting and to increase the performance on a target domain. However, the right composition of data augmentation operations is crucial for learning good representations [87] for multiple domains.

Image Data Augmentation: Image data augmentation is a widely used technique in the field, where the goal is to increase the size and diversity of a training dataset in order to improve the robustness and generalization of DNNs. The process involves applying various transformations to the original images, such as color jittering, random grayscaling, rotation, flipping, and cropping, in order to simulate different real-world conditions. The survey [88] summarizes used data augmentation strategies for image data. Further authors show that the individual image augmentations are class dependent [89]. Whereas some augmentations increase the performance of some classes of a dataset, they harm the performance of others.

Knowledge Graph Data Augmentation: Graph augmentation refers to the process of adding new nodes, edges, or attributes to a graph representation, in order to improve its representational power or generalizability. There are various graph augmentation strategies that have been proposed in the literature, each with its own advantages and limitations. Some of these strategies include node splitting, graph addition, node embedding, and graph convolution. These methods can be used individually or in combination, depending on the specific task and the desired outcome. The survey [90] provides an overview of graph augmentation strategies.

2.4.3 Labels

In DL, i.e. supervise learning, the labeling of data samples plays an important role in the training of ML models. These labels describe how the DNN should interpret the individual image based on the task. For instance, in classification, the label provides the name of the object in the image. Labels can be either hard or soft, and the choice of labeling method affects the performance and interpretability of the model. Hard labels are binary or categorical representations that assign a unique class to each data sample, while soft labels are continuous or probabilistic representations that assign a probability distribution over the classes for each sample.

Hard-Labels: The common approach in supervised learning is to assign a hard-labels to an image. A hard label is used as one-hot encoding to a data sample, e.g. an image is a "cat". Hard labels provide a clear and decisive assignment of class membership, which makes them simple and straightforward to use in the training process. However, a hard label treats all instances of a class in a similar way and does not take into account the fact that instances of one class may be more similar to some other classes than to others.

Soft-Labels: Soft labels are continuous or probabilistic representations of class labels in DL [91]. Soft labels induce additional knowledge about the similarity of the image to other classes. Soft labeling comes with the assumption that assigning a discrete class to an image can be a hard task in the real world. In this labeling method, each data sample is assigned a probability distribution over

the classes. For instance, if the task is image classification, a sample may be assigned a probability distribution such as $[0.6, 0.3]$ over the classes "cat" and "dog", respectively. This means the sample is 60% likely a cat and 40% likely a dog. However, the use of soft labels can also introduce additional uncertainty and noise into the training process.

2.4.4 Network Architecture

Network architectures define or constrain the interaction between specific parameters in the DNN. Therefore, DNN architectures can be seen as hard constraints to the learning problem, such that specific architectures are only suited for specific tasks.

Spatial Dependency: Object recognition tasks rely on a strong inductive bias of spatial localization and translation invariance, such that an object is to be recognized without considering its localization in the image. For instance, CNNs make use of spatial dependency. These networks assume that important features are based on local neighborhoods. Therefore, they learn filter kernels of small size, e.g. 5×5 or 3×3 , and slide them over the whole image. By sharing learned kernels they can save parameters and avoid learning correlations of unrelated far regions in the image.

Sequential Dependency: Sequential data in texts, or temporal data in videos, need to take previous data into consideration. These temporal dependencies can be modeled by recurrent neural networks. Typical representatives are RNNs, LSTMs, and GRUs. Their main principle is to aggregate a state over several timestamps in a dynamic environment. Hereby, it is assumed that every timestep representation is dependent on an aggregation of previous timestep representations.

Structural Dependency: In graph neural networks the architecture is defined by the underlying graph structure. A graphical neighborhood is defined by relationships rather than spatial distances. Every node is dependent on its connected neighbors through its relationships. Especially, for molecule synthesis or social network prediction the provided graph structure helps to constrain the problem space.

2.4.5 Training Objective

The training objective is also known as loss function in DL. Loss functions are mathematical functions that measure the difference between the predicted output of a model and the true output, i.e. label. The choice of a suitable loss function depends on the task at hand, as well as the type of data being used. For image-only methods, the *mean squared error* (MSE) is commonly used for regression problems, while the cross-entropy loss is used for classification problems. There are several derivatives of both basic loss functions. Since the applicable loss function depends on the task and

on the modality being compared, we focus on loss functions being applied for joint embedding spaces of semantic and visual embeddings.

Aligning latent samples in high-dimensional space often leads to problems, as distances are not well defined. Therefore, a common idea is to reduce dimensionality before comparison. A projection head is a linear layer or a few-layer NN that learns to convert a higher-dimensional output into a lower-dimensional output. To use various dimensional KGE with the visual embedding space of the neural network a projection head can be integrated into the training pipeline of the DNN. Since visual and semantic information can be encoded in a vector-based embedding space forming \mathbf{h}_v and \mathbf{h}_s , there are several training objectives to learn a joint embedding. The objectives and also DNNs are optimized using an optimization method. The optimization method minimizes an objective, that measures how far apart the ground truth from the predicted probability distribution or value is. The most common principle to derive specific objectives that are good estimators for different models is the maximum likelihood principle. These objectives can be seen as cross entropy between the empirical distribution of the training set and the probability distribution defined by model [48].

Here we present some of the basic objectives used in visual transfer learning using KG, which can be augmented with additional regularization terms or hyperparameters. There are configurations of visual and semantic embedding space that only allow certain objectives to be applied. We define $\mathbf{l} \in \mathbb{R}^K$ as the network’s output vector (logits), and $\mathbf{t} \in 0, 1^K$ as the one-hot encoded vector of targets, where $\|\mathbf{t}\|_1 = 1$. We refer to visual data as x_v and semantic data as x_s , and equally to visual embedding as \mathbf{h}_v and semantic embedding as \mathbf{h}_s .

The Effect on Network Layers: Recently there was the idea that a loss function can influence how the DNN learns from data. However, work [92, 93] showed that the objectives have a smaller impact on the learned DNN than suspected. A loss function only influences the penultimate layers of a DNN [94]. Using centered kernel alignment to measure similarity between hidden representations of networks, they find that differences between loss functions are only visible in the last layers of the network. They look more closely at the penultimate layer representations and find that different objectives and hyperparameter combinations lead to dramatically different levels of class separation.

Pointwise Objectives

Softmax Cross-Entropy (CE) [95]: CE is the most common objective to learn multi-class classification tasks. The softmax represents a probability distribution over a discrete variable with K possible values, i.e. classes. CE learns the DNN end-to-end by comparing the logits \mathbf{l} with the

target vector \mathbf{t} and is given by

$$L_{CE}(\mathbf{l}, \mathbf{t}) = - \sum_{k=1}^K t_k \log \left(\frac{\exp(l_k)}{\sum_{j=1}^K \exp(l_j)} \right) \quad (2.4)$$

$$= - \sum_{k=1}^K t_k l_k + \log \sum_{k=1}^K \exp(l_k). \quad (2.5)$$

Mean Squared Error (MSE): MSE is the most intuitive way of attracting two vectors and is given by

$$L_{MSE} = \frac{1}{K} \sum_{k=1}^K \|\mathbf{h}_{s,k} - \mathbf{h}_{v,k}\|^2. \quad (2.6)$$

The MSE loss calculates the Euclidean distance and maps a training image $x_{v,k}$ and its visual feature vector $h_{v,k}$ to a semantic embedding vector $h_{s,k}$, corresponding to the same class k [96].

However, using the Euclidean distance as a metric fails in high-dimensional space [97]. An alternative metric in high dimensions is the cosine distance, which is given by $sim(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} / \|\mathbf{u}\| \|\mathbf{v}\|$.

Pairwise Objectives

Pairwise objectives [98] always rely on the information of positive and negative samples. They have the goal to pull positive visual embedding vectors $\mathbf{h}_{v,p}$ to its corresponding semantic embedding anchor vector $\mathbf{h}_{s,a}$ and push negatives $\mathbf{h}_{v,n}$ away [99].

Triplet and Hinge Rank Loss [100]: The triplet and hinge rank loss requires explicit negative sampling. It uses a margin α as a regularization term and it is given by

$$L_{tri} = \sum_{n \neq p} \max[0, \alpha - sim(\mathbf{h}_{s,a}, \mathbf{h}_{v,p}) + sim(\mathbf{h}_{s,a}, \mathbf{h}_{v,n})]. \quad (2.7)$$

Contrastive Loss: The contrastive loss extends the triplet loss by a version of the softmax and handles multiple positives and negatives at a time and is given by

$$L_{con} = - \log \frac{\exp(sim(\mathbf{h}_{s,a}, \mathbf{h}_{v,p})/\tau)}{\sum_{n=1}^{2N} \mathbb{1}_{n \neq a} \exp(sim(\mathbf{h}_{s,a}, \mathbf{h}_{v,n})/\tau)}. \quad (2.8)$$

where, $\mathbb{1}_{n \neq a} \in \{0, 1\}$ is an indicator function that returns 1 iff $n \neq a$, and $\tau > 0$ denotes a temperature parameter.

2.4.6 Optimization Method

The basic principle of a DNN is to adjust many parameters to map an input, i.e. image pixels, to a desired output. The process of finding the optimal set of parameters that satisfy the mapping

task is called optimization. Optimization involves searching through a high-dimensional space of possible parameters, e.g. weights of the DNN, to find the optimal set that produces the lowest loss between the output of the DNN and the labels. In the context of DNNs, gradient-based methods are typically used for optimization, where each model parameter is assigned a gradient with respect to the gradient of the loss function. The individual parameter gradient provides information on how to update the parameter efficiently to minimize the overall loss.

One of the most widely used gradient-based optimization methods for DNNs is *stochastic gradient descent* (SGD) [101]. SGD is a version of gradient descent that computes the gradient of the loss function on a small batch of data at a time, rather than the entire dataset. This approach is computationally more efficient and allows for fast convergence to a good solution. However, SGD has some limitations such as the tendency to get stuck in local minima. In recent years, several extensions to SGD have been developed to improve the learning of DNNs. Some of these methods include AdaGrad, Adam, and RMSProp [102]. These optimization methods incorporate adaptive learning rates, momentum, and other advanced techniques to help the model to converge to an optimal solution. However, the choice of an appropriate optimization method depends on the characteristics of the problem being solved, as well as on other inductive biases of the DNN, such as the architecture or the distribution of the dataset.

Nearly all variants of SGD rely on back-propagation [103] as a learning algorithm to distribute the gradient of the loss to the individual parameters of the DNN, based on their influence on the final outcome. Iteratively, every parameter is assigned with a gradient and then updated in the direction to minimize the final error. It is worth noting that the optimization method heavily influences the way a DNN learns from data and that there are more and more conjectures that the human brain learns differently from data than DNNs. For example, the brain can learn from a few training examples with many neurons, while DNNs learn dense representations into a few neurons from many training examples. Moreover, there is no convincing evidence that the cortex explicitly propagates error derivatives or stores neural activities for use in a subsequent backward pass [104]. However, research on alternative learning algorithms is still in its early stages and there are not many viable alternatives that scale yet. Alternative learning algorithms are for example the Boltzmann Learning Algorithm [105] or the Forward-Forward Algorithm [104].

However, not only the optimization method itself but also the choice of hyperparameters and the initialization of the weights can have a significant impact on the performance of DNNs. Therefore, it is important to tune these parameters carefully to achieve the best possible performance.

Hyperparameter Selection: Hyperparameters are parameters that are set before training and are not learned during training. These include variables such as the number of layers, the number of neurons in each layer, the learning rate, and the stack size. Choosing appropriate hyperparameters is critical for a good performance of DNNs. For example, if the learning rate is set too high, the model may overshoot the optimal solution and oscillate, causing it to miss the optimal solution. On

the other hand, if the learning rate is set too low, the training process may get stuck in local minima and is relatively slow. Similarly, if the DNN is too small, the DNN may not be able to capture complex patterns in the data, while if it is too large, the DNN may overfit.

However, selecting the right hyperparameters remains a challenge, and most approaches use heuristics that try different hyperparameter configurations and select the best one. A common approach to hyperparameter selection is grid search, where the user specifies a range of values for each hyperparameter and the system trains the model for each combination of values in that range. Grid search is excellent for randomly testing combinations that are known to generally perform well. Another approach is random search [106], in which the system randomly selects values from a given range for each hyperparameter and trains the model for each combination of values. Random search is excellent for discovering hyperparameter combinations that would not have been thought of intuitively, although it often takes more time to execute. Sometimes more advanced methods are used, such as Bayesian optimization and evolutionary optimization.

Weight Initialization: Weights are usually initialized randomly before the training process begins. The initial values of the weights can affect how fast the model converges and whether it gets stuck at a local minimum. If the initial weights are too large, the model can go into saturation and produce very small gradients, slowing down the learning process. On the other hand, if the initial weights are too small, it may take a long time for the model to converge or get stuck in local minima.

Therefore, choosing the right method to initialize the weights is important to achieve fast convergence of the model and better performance. A common approach for initializing the weights is to take the initial values from a normal distribution with a mean of 0 and a standard deviation of 1. However, this approach can lead to slow convergence and poor performance in some cases.

A better approach is to use initialization techniques specific to the activation function used in each layer. For example, the Xavier initialization method [107] scales the weights by the square root of the number of inputs to the layer, which has been shown to improve convergence and reduce the likelihood of disappearing or exploding gradients. However, the Xavier initialization method has been found to cause problems when initializing DNNs that use the *rectified linear unit* (ReLU) activation function. Another popular initialization method that can handle ReLU activations is the He initialization method [108], which scales the weights using the square root of the number of neurons in the previous layer.

Chapter 3

Transfer Learning using Prior Knowledge

Transfer learning is a powerful approach to learning that enables the application of knowledge gained from one domain to another. The concept of transfer learning may have originated in educational psychology. Psychology assumes that people frequently engage in transfer learning in everyday life. They may reuse learned knowledge to better interpret new things or to learn them more quickly.

Therefore, there are also attempts to transfer the theory of transfer learning to DL. Since training DNNs is a rather tedious task and requires a lot of data and computational power, the idea of training a model on a source domain and reusing it for related target domains with sparse data is quite prominent in DL. Moreover, the principle of DL, where features are learned in a training domain to be applied in a test domain, implicitly requires transfer learning skills since reality is always different from the training data [109].

While DNNs tend to be highly specialized in the domain on which they were trained, this specialization may come at the cost of poor generalization to new and unknown domains [110]. Therefore, to achieve optimal performance in transfer learning for DL, it is important to strike a balance between specialization and generalization. To this end, DL approaches either try to avoid transfer learning by increasing the source domain size or focus on inducing additional biases in the learning process. The main goal is always to relax the strong training data dependency and provide additional knowledge that is valid for both source and associated target domains.

However, finding the right inductive bias for a transfer learning task is not straightforward, as the choice hardly depends on the setting, e.g. the data, the domain shift, or the model. Moreover, these inductive biases are inflexible, limited, and do not correspond to the way humans think about domains. Since humans are well suited to provide additional high-level knowledge about the relationship between the source and target domains, there is an urgent need to find such an interface

that can better utilize human prior knowledge.

We hypothesize that symbolic prior knowledge about a domain is much more flexible and better suited to describe human prior knowledge and thus provide DL with additional knowledge. In this context, we introduce the notion of transfer learning using prior knowledge, where DNNs are supported with prior knowledge to decide what distinguishes a good feature from an incorrect one. Additionally, we argue that such prior knowledge can be well encoded and formalized with KGs since KGs provide the necessary tools to encode such knowledge in an explicit and implicit format. Therefore, transfer learning using KG can help to learn better-generalized DNNs and transfer knowledge across different domains and tasks. It has the potential to enable more effective and robust use of DL for many real-world applications.

In this chapter, we address the following research question:

RQ1: Why can prior knowledge encoded in a KG improve DL-based visual transfer learning?

The main contribution of this chapter is summarized as follows:

- *Structured analysis of why prior knowledge encoded in a KG can improve DL-based visual transfer learning.*

Most of the topics described in this chapter are already published in:

- **Sebastian Monka**, Lavdim Halilaj, and Achim Rettinger. 2022. A survey on visual transfer learning using knowledge graphs. *Semantic Web* 13, 3 (2022), 477–510.

The chapter introduces the concept of transfer learning using prior knowledge. Therefore, it starts in Section 3.1 with relevant fundamentals of transfer learning, explaining transfer learning scenarios, such as input domain change and output domain change. Further, it continues with Section 3.2, where the problem of transfer learning is related to DL. Therefore, the trade-off between specialization and generalization is explained and methods for enabling DL for transfer learning are discussed. Section 3.3 then introduces transfer learning using KGs. Hereby, we introduce the notion of explicit and implicit prior knowledge. Further, we show that KGs are well suited to encode both types of knowledge and therefore are well suited as a representation format for prior knowledge. Finally, Section 3.4 provides an overview of datasets for visual transfer learning using KG, including generic KGs and visual transfer learning tasks with and without prior knowledge.

3.1 Fundamentals of Transfer Learning

Transfer learning is a phenomenon that can be observed when humans interpret or learn new things, e.g. languages, music instruments, or sports. If the new task is related to a task that has been

already acquired in the past, learned knowledge can be reused. However, transfer learning is not always beneficial and can be separated into positive and negative transfer. Positive transfer refers to a scenario where acquired knowledge of a source domain can be leveraged to better understand a novel target domain. Negative transfer refers to a scenario where acquired knowledge of a source domain hinders understanding a novel target domain [111]. Whether positive or negative transfer will occur may depend on several factors, such as the relevance between the source and the target domains and the learner’s capacity of finding the transferable and beneficial part of the knowledge across domains. [112]. The main goal of transfer learning can be directly transferred to DL and is therefore to enable positive and avoid negative transfer.

A theoretical framework for transfer learning is presented as follows [113, 114]:

Given a source domain D_S with input data X_S , a corresponding source task T_S with labels Y_S , as well as a target domain D_T with input data X_T and a target task T_T with labels Y_T , the objective of transfer learning is to learn the target conditional probability distribution $P_T(Y_T|X_T)$ with the information gained from D_S and T_S where $D_S \neq D_T$ or $T_S \neq T_T$.

3.1.1 Transfer Learning Scenarios

Transfer learning can occur in two distinct scenarios, either input domain change or output domain change. Input domain change refers to scenarios, where the same type of object is represented using different but related types of input. Output domain change happens if a new type of object that is related to the initial type of object is added to the possible outputs.

Figure 3.1 describes both transfer learning scenarios: a) input domain change, where a distribution shift is happening in the input, and b) output domain change, where a distribution shift is happening in the output.

Input Domain Change: Input domain change refers to scenarios where the initial input changes from source to target domain. For instance, the source domain, on which the DNN was trained, contains images of road signs, and the target domain, to which the DNN is applied, contains the same road sign images deviating in terms of weather, artificial changes, or country.

The scenario of input domain change includes two distinct tasks, domain generalization, and domain adaptation. *Domain generalization* is a transfer learning task with access to labeled source domain data and unlabeled target domain data. Domain generalization aims to extract implicit knowledge of the source domain D_S and transfer this knowledge to an unknown target domain D_T [115, 116]. If domain generalization has access to an additional set of labeled target data X_T , the task is called *domain adaptation*. Figure 3.1 a) shows a *stop* road sign of a source domain and various variations of the *stop* road sign of the target domain. Examples of the target domains are the same road signs with deviating weather, artificial changes, or different countries.

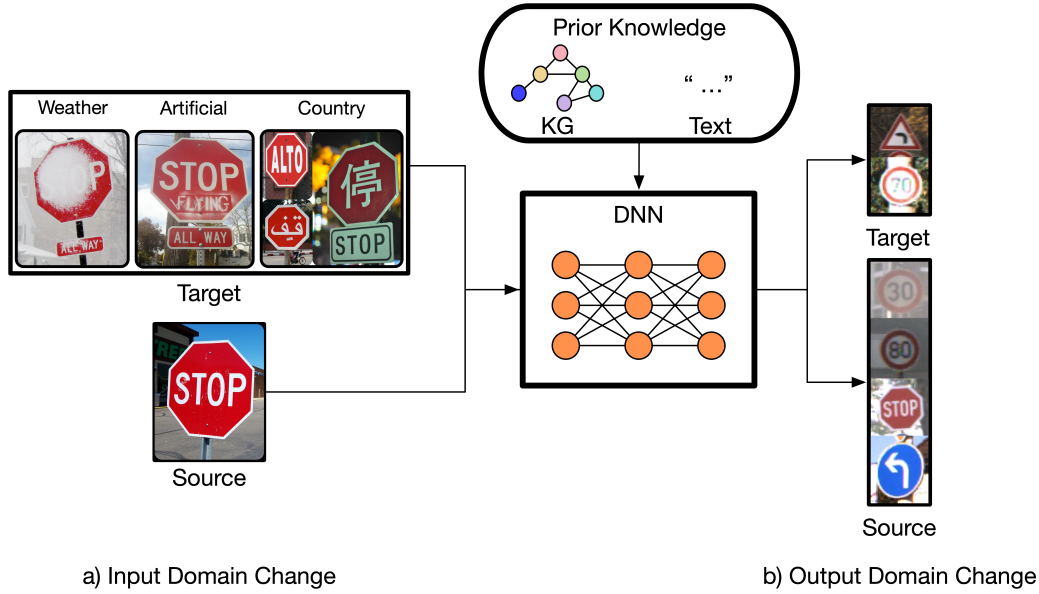


Figure 3.1: Transfer learning can be divided into two distinct scenarios: a) input domain change, where a distribution shift is happening in the input, and b) output domain change, where a distribution shift is happening in the output. Prior knowledge, such as knowledge graphs or textual descriptions, can help a DNN deal with these scenarios.

Output Domain Change: Output domain change refers to scenarios where the initial output changes from the source to the target domain. For example, the DNN was trained on road signs, including *speed limit 30*. When applied to a target domain, the model must also predict a *speed limit 70* that was not present in the initial domain and therefore not present at training time.

The scenario of output domain change includes two tasks, zero-shot learning, and few-shot learning. *Zero-shot learning* is a transfer learning task with labeled source domain data and unlabeled target domain data. Zero-shot learning aims to extract implicit knowledge of the classes in the source domain task T_S and transfers this knowledge to unknown classes of the target domain task T_T [113]. If zero-shot learning has access to an additional set of labeled target data X_T , the task is called *few-shot learning*. Figure 3.1 b) depicts road signs of a source domain and additional road signs of a target domain. The goal of the tasks is to leverage the knowledge of the source domain road signs and transfer it to the ones from the target domain.

3.1.2 Strategies in Transfer Learning

It is an open question of how to effectively deal with transfer learning scenarios in DL. Since the strong dependency on the training data is a fundamental problem of DL, there is no satisfactory solution to deal with distribution shifts so far.

DL pipelines that deal with distribution shifts can be grouped into three main categories as

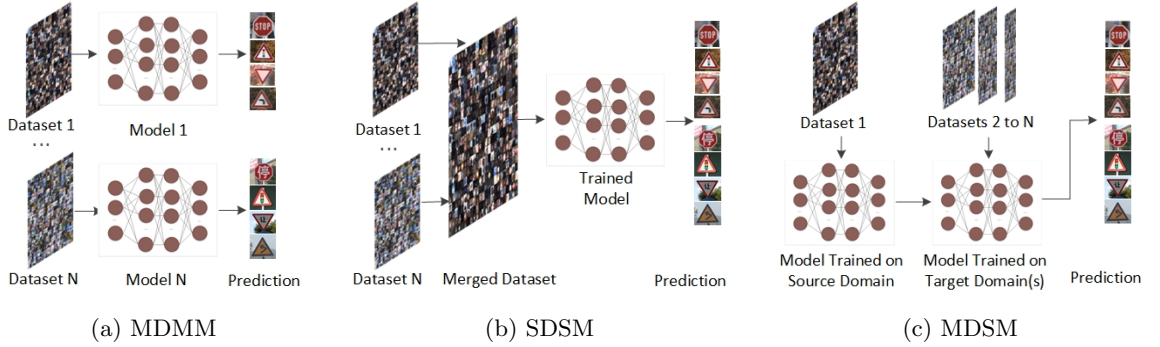


Figure 3.2: Strategies for handling transfer learning scenarios in DL. Approaches of DL that deal with multiple domains can be categorized into a) *Multiple Domains Multiple Models* (MDMM); b) *Single Domain Single Model* (SDSM); and c) *Multiple Domains Single Model* (MDSM).

depicted in Figure 3.2: a) *Multiple Domains Multiple Models* (MDMM); b) *Single Domain Single Model* (SDSM); and c) *Multiple Domains Single Model* (MDSM). MDMM approaches treat all datasets as independent and train a respective model for each of them. Therefore, these approaches are very costly to train, and learned knowledge cannot be transferred between datasets. SDSM approaches train a single model on a large dataset merged from many smaller ones. However, it is difficult to create a balanced dataset required by the DNN to learn a general representation suitable for all domains. MDSM approaches train a single model on various datasets at different stages, and can therefore transfer learned knowledge to new domains. However, if trained with the standard cross entropy these models suffer from an unpredictable and error-prone knowledge transfer and catastrophic forgetting, where learned knowledge from previous datasets tends to be forgotten after training on the current dataset.

All approaches deal with the transfer learning problem in different ways. MDMM avoids the transfer learning problem by design, SDSM increases the source domain to also include the target domain, and MDSM needs to find relevant features that also suit the source domain. In the context of this work, transfer learning methods that need to generalize and learn transferable features are methods that belong to MDSM.

3.2 Domain Change in Deep Learning

DL as an ML technique is broadly used to successfully solve CV tasks. Its main strength lies in its ability to find complex underlying features in a given set of images. A common method for training a DNN is to minimize the *cross-entropy* (CE) loss on a training domain, which is equivalent to maximizing the negative log-likelihood between the empirical distribution of the training set and the probability distribution defined by the model. This idea relies on the *independent and identically distributed* (i.i.d.) assumptions as underlying rules of basic ML. This assumption states that the

examples in each dataset are independent of each other, that train and test sets are identically distributed, and that they are drawn from the same probability distribution [48]. However, if the train and test domains follow different distributions, which means that there is a domain change, the i.i.d. assumptions are violated, and DL leads to unpredictable and poor results [117]. This has been demonstrated by using adversarially constructed examples [118] or variations in the test images such as noise, blur, and JPEG compression [119]. Moreover, it is shown [120] that the real world also contains unpredictable distribution shifts to a training and testing dataset, which can lead to critical errors in standard DNNs.

3.2.1 Specialization-Generalization Trade-off in DL

DNNs are designed to learn complex features from large datasets. This is achieved by learning complex patterns and relationships between the input and output data. The main assumption in learning from data is that any learning algorithm will generalize better on some distributions and worse on others [82]. Therefore, every learning process involves a delicate trade-off between specialization and generalization [121].

Specialization refers to the model's ability to fit the training data accurately. Therefore the model is trained to learn every individual feature that maximizes the prediction accuracy of a model on a specific task. However, if a model is too specialized, it can overfit the training data, capturing noise and random fluctuations in the data instead of the underlying patterns. Such features are also called spurious correlations of the training data. When learning these spurious correlations the model will perform well on the training dataset but poorly on the target data. The model memorized the training dataset rather than learning generalizable features. Moreover, If the model can only fit the training data, it will not be useful in real-world applications, since the real world will always deviate slightly from the training data.

Generalization is the ability of a DNN to perform well on new and unseen data. Generalization is critical because it allows the model to be useful in real-world applications. The ability to generalize well is achieved by learning the underlying patterns and relationships that are common across the training data. However, if a model is too generalized, it underfits the training data and does not capture the relevant features in the data. This results in a model that performs poorly on both source and target data, as it is unable to learn the relevant features and relationships in the data.

Therefore, the specialization-generalization trade-off is a fundamental concept in DL and also known as optimization-generalization dilemma [121]. DNNs can either specialize in fitting to the training data, leading to high accuracy on the training data but poor generalization to new data, or generalize better to new data but sacrifice accuracy on the training data. The goal of DL is to find a balance between specialization and generalization that maximizes performance on both the source and target domains. Achieving this balance is critical for the success of DNNs, as it determines their ability to learn from data and make accurate predictions in real-world applications.

3.2.2 Enabling Deep Learning to Generalize

The use of DL to handle dynamic, and changing conditions of the real world has limitations. To enable a DNN to handle such conditions it needs to be able to generalize or transfer to novel domains. Therefore, most approaches either increase the source domain to avoid transfer learning or incorporate inductive biases, such as those introduced in Section 2.4, into a DNN to handle the specialization-generalization trade-off.

Scaling the Source Domain: Enabling DL to generalize to novel domains is a quite challenging task due to the specialization-generalization trade-off. Therefore, one approach to make DNNs generalizable is scaling the source domain, such that a target domain will already be inside the source domain and therefore be seen at training time. As described in Section 3.1.2, we assign such models to the category SDSM. Having such a huge source domain avoids the difficult task of transfer learning in some sense. Recent methods that use that trick are for example foundation models [122] of language and vision. Through a self-supervised training task, they enable large-scale training on a huge amount of data from the Internet.

However, whereas this approach seems to provide good transfer learning results, it is quite debated if such models are leading to better transferable features. One assumption is that these models learn better-generalized features through a self-regularization due to a large amount of training data and its large variability. Another assumption is that these large models just remember the training data and that the huge domain is just a way to avoid the transfer learning problem of DL since the model has already seen data from the target domain. Therefore, it needs to be further investigated if scale can improve the transfer learning capabilities of DL.

Usage of Inductive Biases: To enable models of the category MDSM, as described in Section 3.1.2, to transfer, the learning process from unstructured data needs to be influenced. Consistent with the specialization-generalization trade-off, there are many possible solutions to a learning problem that exhibit equally performance on the training data. Given a finite training set, the only way to generalize to new input configurations is then to rely on some assumptions about the preferred solution [83]. These assumptions that guide the learning process are called inductive biases.

Inductive biases are non-learnable parts of the DNN that cause the learning algorithm to favor solutions with certain properties. In Section 2.4 we provided an overview of the inductive biases of DL. Inductive biases occur in various parts of the DNN, such as in the input data, augmentation methods, architecture, training objectives, labels, and optimization methods. They include assumptions such as function smoothness, translation invariance for CNNs, and a hierarchical organization of concepts as a general assumption of DL [4].

Inductive biases are mostly chosen by human intuition. For instance, if the target domain consists of just grayscale images, it is helpful to provide grayscale data augmentation to the source domain

data when learning the DNN that should be applied to the target domain. Another strong inductive bias for vision is the usage of convolutional layers in CNNs, which assume for instance translation invariance such that objects and features retain their semantic meaning no matter in which part of the image they appear. Using such inductive biases DL can learn more efficient and more robust models in the real world. An important question for AI research aiming at human-level performance then is to identify relevant inductive biases for the real world. However, there is still much work to be done to improve our understanding of inductive biases and how to incorporate them into DL [83].

In summary, incorporating inductive biases into DL can help to enable transferability and generalizability. Inductive biases are parts of DNNs that are human-defined and not learned. They are highly dependent on human experts and their prior knowledge and can be seen as constraints on how DL learns from unstructured data. However, there is still much work to be done to improve our understanding and incorporate these biases effectively. Achieving a balance between specialization and generalization requires careful consideration of inductive biases and is essential for the success of DNNs in real-world applications.

3.3 Transfer Learning using Knowledge Graphs

Transfer learning using KG relates to the task of improving the transfer learning capabilities of DL using prior knowledge encoded in a KG. We introduce the effect of prior knowledge on human cognition and relate important concepts to the development of AI systems and transfer learning. Further, we show that prior knowledge can be divided into explicit and implicit knowledge and explain the different types in detail.

To use prior knowledge in combination with DL we suggest using KGs as a representation format and describe why KGs are well suited to encode prior knowledge for transfer learning. Hereby, we argue that KGs are ideal for representing prior knowledge since they are able to encode both, explicit and implicit knowledge. Therefore KGs can suit as an interface to explicit human knowledge, as well as to implicit embeddings learned from data through a KGE.

The theoretical framework for transfer learning using KG can be defined as follows:

Given a source domain D_S , a target domain D_T , and their corresponding tasks T_S and T_T as well as the prior knowledge represented by a KG and embedded into a latent space \mathbf{h}_s , the objective of a transfer learning task using a KG is to learn a probability distribution $P_{KG}(Y_T|X_T, KG)$ with the information gained from D_S and T_S where $D_S \neq D_T$ or $T_S \neq T_T$ and the prior knowledge given by KG.

3.3.1 Types of Prior Knowledge

The concept of prior knowledge in human cognition has been used in various fields of study, such as psychology, education, and cognitive science, among others. Prior knowledge refers to knowledge that an individual has acquired before encountering a new problem or situation. It is the knowledge and experience that a person brings to a new situation that can be used to aid in learning or problem-solving. Prior knowledge has typically been shown to facilitate learning [123], increase the rate at which novel categories are learned [124], decrease prediction errors during learning, and make it possible for learners to acquire categories with a complex relational structure [125]. Prior knowledge can play an important role in learning and problem-solving by providing a foundation on which new knowledge can be built [125, 126] It is assumed that prior knowledge interacts with other variables to influence learning outcomes [127].

In the context of AI and DL, prior knowledge also refers to any knowledge that exists about the problem domain before the learning process begins [4]. In this context, the term prior knowledge is also commonly used to refer to knowledge that can be leveraged to improve the performance of DL. Consistent with theories of human cognition, prior knowledge is not only important to solve transfer learning problems, but also a way to influence the way a DNN learns from unstructured data [128].

As seen in Section 3.2, most transfer learning approaches from DL address the domain shift problem by reducing dependence on source domain data through inductive biases. As shown in Section 3.2.2 inductive biases can be seen as a specific type of prior knowledge, since they influence what kind of features are learned on the training data. However, choosing the right inductive bias for a problem is a complicated and error-prone process [83]. Therefore, DL approaches also incorporate other types of prior knowledge to solve the transfer learning issue of DL. Prior knowledge can be divided into explicit and implicit knowledge [129].

Explicit Knowledge: Explicit knowledge refers to knowledge that is codified, recorded, or otherwise formalized in some tangible form that can be easily transmitted and communicated to others. It is the knowledge that can be articulated and expressed through language or other symbols, and is often easily documented, stored, and shared. We refer to explicit knowledge as a form of symbolic knowledge, that is represented and processed using symbols, such as words, numbers, or other abstractions. Explicit knowledge can be constructed, is intuitive, compositional, and controllable by human experts. Examples of explicit knowledge include information from textbooks, technical specifications, scientific articles, and other written or visual materials that convey information in a structured and systematic way often represented in graphs.

Implicit Knowledge: Implicit or tacit knowledge is often deeply ingrained in an individual's cognitive and experiential background and can be difficult to communicate to others through formal means. It is difficult to articulate or codify and is often based on personal experience, intuition,

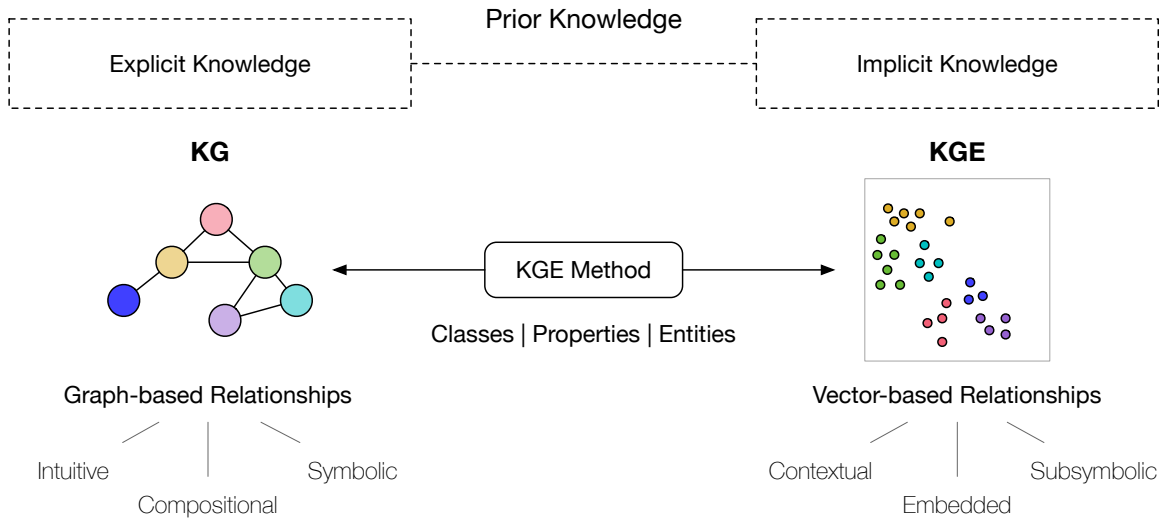


Figure 3.3: A *knowledge graph* (KG) can represent prior knowledge explicitly and implicitly. This makes a KG an ideal representation format for prior knowledge, as it interfaces with explicit human knowledge through its *knowledge graph embedding* (KGE) and connects to implicit embeddings of data from other modalities, such as vision or language.

and informal interactions. Implicit knowledge is therefore embedded knowledge based on a specific context. As shown in Section 2.3 it can be extracted from data and stored using feature extractors, forming embeddings of different modalities using specific contexts. These embeddings are a dense representation of contextual knowledge illustrated through vectors in high-dimensional space. Therefore, implicit knowledge is able to encode a large amount of information in an efficient way.

3.3.2 Knowledge Graph as a Representation Format

As introduced in Section 2.2.3, KGs are an effective representation format for capturing prior knowledge. Through the development of KGE methods, they are able to represent both implicit and explicit types of knowledge as depicted in Figure 3.3. To encode explicit prior knowledge, KGs use a graph-based representation format, where classes, properties, and entities are represented through nodes and their relationships through edges. These graphs allow for interaction with human experts and include underlying concepts to unify data. Due to their established tool set, KGs provide excellent preconditions for modeling, reusing, reasoning, and interlinking all sorts of heterogeneous knowledge. KGs can include domain knowledge, from human experts, taxonomies, books, or other sources of data, which can be used to model classes, entities, their properties, and relationships. Since KGs offer a natural interface for encoding, modifying, and using explicit prior knowledge, they allow for a more comprehensive understanding of the domain and enable effective knowledge management.

Moreover, KGs are also able to represent their knowledge using an implicit format. With the recent development of KGE methods that transform KGs into vector spaces, the symbolic encoded prior knowledge of a KG can be transformed into an implicit representation. Therefore KGEs can be linked to implicit embeddings of data from other modalities, such as vision or language. In contrast to visual or language embeddings that are based on the collection of the underlying training data, KGE methods provide more flexibility to manually influence how the training data, i.e., the KG, is built and how the implicit knowledge is formed. As shown in Section 2.3.2, there are many different KGE methods that influence how knowledge is represented in its implicit form. Implicit knowledge of a KG, i.e. KGE, is therefore a form of knowledge represented in vector space that reflects similarities between different entities and classes based on their positions in an embedding space. Hereby, KGEs encode the knowledge about relationships of objects due to the similarities and distances in the embedding space.

In summary, A KG can represent prior knowledge explicitly and implicitly. This makes a KG an ideal representation format for prior knowledge, as it interfaces with explicit human knowledge through the KG and connects to implicit embeddings of data from other modalities, such as vision or language, through its KGE. KGs are flexible, interpretable, and modifiable, and come with an established toolset to build, reason, interact, and embedding heterogeneous sources of information. Therefore, KGs can be considered an ideal representational format for encoding prior knowledge for visual transfer learning.

3.4 Resources for Visual Transfer Learning using KG

The field of visual transfer learning using KG is fairly new, and only a few datasets exist that can be used out of the box. Therefore, many visual transfer learning approaches use datasets that just contain a type of prior knowledge, such as attributes or descriptions. This prior knowledge can be transformed into a KG representation. Since building meaningful KGs from scratch can be a challenging task and KGs are designed on the principle of reusability, it is possible to reuse or enrich generic KGs also for related tasks [41].

Therefore, this section first provides an overview of generic KGs that can be used with prior knowledge in Table 3.1. Second, we provide in Table 3.2 a list of datasets for knowledge-based ML and visual transfer learning. We categorize these datasets based on their modality of knowledge into: a) Attribute-augmented image datasets with textual image or class attribute descriptions; b) Language-augmented image datasets, providing additional textual descriptions of the images; c) Knowledge graph-augmented image datasets, containing metadata of class relations in a KG; d) Image datasets without prior knowledge, used for zero-shot learning and domain generalization tasks.

	WordNet	DBpedia	Wikidata	ConceptNet	CSKG
#triples	2,6M	1,1B	14,5B	34M	4,6M
describes	words, concepts	instances, concepts	instances, concepts	objects, actions, relations	objects, actions, relations
type representation	thesaurus directed	encyclopedia hypergraph	encyclopedia hyper-rel	commonsense directed	commonsense hyper-rel
external sources	-	-	-	WordNet DBpedia OpenCyc Wiktionary	WordNet Wikidata ConceptNet
generation	manual	crowdsourced	crowdsourced	crowdsourced	automatic
ZSL use cases	[130, 131, 132] [133, 134, 135]	[136]	-	[137]	-
DG use cases	[138, 139, 140] [141, 142, 143] [144]	-	-	[145]	-
VQA use cases	-	[146]	-	-	-
release date	1985	2007	2012	2017	2021
version	2011-06 (3.1)	2022-09	2023-01	2020-05 (5.8)	2020-12
example	lion.n.01 hypernymy	dbr:Lion dbp:description	wd:Q140 wdt:P279	/c/en/lion/n /r/RelatedTo	/c/en/lion/n /r/RelatedTo

Table 3.1: We provide an overview of available generic KGs and compare them based on distinct properties, e.g. size, content, and representation type. We also present relevant work from *domain generalization* (DG), *zero-shot learning* (ZSL), and *visual question answering* (VQA), which already use the KGs.

3.4.1 Generic Knowledge Graphs

Over the years, several open-access KGs have been created by various community initiatives. These graphs contain universal knowledge which potentially can be used as prior knowledge in various scenarios since KGs can be interlinked or reused for related tasks. Generic KGs differ in terms of size, content, and representation. Their size ranges between thousands and billions, their content varies from domain-specific to commonsense knowledge, and their representation is either manually curated, crowdsourced or automatically extracted from the web.

In the following, we compare some of the most common generic KGs summarized in Table 3.1. However, for deeper insights and more KGs, we refer to the survey of Färber et al. [147].

WordNet [148]: Wordnet is a lexical database of English words, which organizes its words into synonym sets and organizes these sets into a hierarchy. Since WordNet 3.1 is frequently used in visual transfer learning scenarios with prior knowledge, is available in RDF, and encodes rich semantic concepts, we define it as a generic KG for the scope of this work. WordNet, first released

in 1995, is an online lexical reference system for English nouns, verbs, and adjectives which are organized into *synonym sets* (synsets), each representing one underlying lexical concept. WordNet superficially resembles a thesaurus, in that it groups words based on their meanings. There are 117 thousand synsets, each synset is linked with other synsets by super-subordinate relations, forming a hierarchical structure of instances, concepts, and categories whereas all are linked with the root node, *entity*. Since WordNet is manually built, it is the most structured and accurate, but also the smallest and less detailed KG.

DBPedia [149]: DBPedia is a structured data version of Wikipedia. It is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against datasets derived from Wikipedia and to link other datasets on the Web to Wikipedia data. DBPedia answers questions like "Give me the birthplace of Albert Einstein?" or "Give me all movies Leonardo Di Caprio has starred in?". The underlying structure of DBpedia is a hypergraph model in Section 2.2.3 where facts are represented via binary and n-ary relationships. The English version of the DBpedia knowledge base describes 4,58 million things, out of which 4,22 million are classified in a consistent ontology, including 1,445 thousand persons, 735 thousand places, and 411 thousand creative works ¹. However, DBPedia is known to be messier than other generic KGs, where the vocabulary only consists of titles of Wikipedia articles.

Wikidata [150]: Wikidata is a KG, built collaboratively by humans or automated agents. It encapsulates facts about the world entities organized in a form of complex statements. The basic structure comprises items defined with a label and several aliases. In addition, Wikidata contains some sense of basic commonsense knowledge [151] which allows for performing several sophisticated downstream tasks based on reasoning capabilities. The facts within Wikidata are represented as a hyper-relation graph in Section 2.2.3, where relations are enriched with additional information known as qualifiers [79]. These qualifiers enable the disambiguation of complex facts about the same entities in different contexts. Currently, Wikidata has about 101 million items, e.g. 23 million scholarly articles, six million of them are humans, five million astronomical objects, and two million administrative entities ².

ConceptNet [152]: The ConceptNet graph encodes commonsense knowledge about everyday objects. It is a freely-available semantic network, designed to help computers understand the meanings of words that people use. ConceptNet originated from the crowdsourcing project Open Mind Common Sense, which was launched in 1999 at the MIT Media Lab. Its knowledge is collected from many sources including resources created by experts, crowdsourcing, and games with a purpose.

¹<https://wiki.dbpedia.org/about>, accessed on 27 January 2023

²<https://www.wikidata.org/wiki/Wikidata:Statistics>, accessed on 27 January 2023

ConceptNet connects words and phrases of natural language. It was initially designed to represent the general knowledge involved in understanding language, improving natural language applications by allowing the application to better understand the meanings behind the words people use. Information within ConceptNet is modeled as a directed labeled graph in Section 2.2.3, where concepts are connected via binary relationships. It contains approximately 34 million facts ³.

CSKG [153]: CSKG is a commonsense knowledge graph initially designed to use heterogeneous sources for question answering and natural language inference tasks. The KG unifies seven generic KGs into a single representation. It integrates the commonsense source ConceptNet, the visual commonsense source Visual Genome, the procedural source ATOMIC, the general-domain source Wikidata, and three lexical sources, WordNet, Roget, and FrameNet. CSKG is represented as a hyper-relational graph 2.2.3, by using the KGTK data model. Moreover, for CSKG text and graph embeddings are provided, showing that CSKG is well-connected and that its embeddings provide a useful entry point to the graph. The KG can be used for generalizable downstream reasoning and for pre-training of language models and is publicly available.

3.4.2 Image Datasets

Image datasets are used to train and test models for various CV tasks, such as image classification, object detection, or semantic segmentation. To enable the models to generalize many datasets provide additional information to the images, e.g. attributes or descriptions. In the scope of this work, we refer to that additional data as prior knowledge. While some datasets build on the structures of earlier knowledge bases, such as ImageNet on WordNet, others have additional attributes or descriptions for their images, such as AwA or MS-COCO.

We provide a categorization of visual datasets in Table 3.2. We categorize the datasets in terms of their modalities, their task, the type of prior knowledge they provide, and their release date.

Attribute-Augmented Image Datasets: We define attribute-augmented image datasets as image datasets with additional textual image and class attributes. In particular, every image contains attributes that describe the context in or even beyond the visual information.

AwA [154]: The *Animals with Attributes* dataset consists of over 30 thousand images with pre-computed reference features for 50 animal classes, for which a semantic attribute annotation is available from studies in cognitive science. However, as the AWA images do not have a public copyright license, only some computed image features, i.e. SIFT [51], DECAF [155], VGG19 [156] of the AwA dataset are publicly available, rather than the raw images. Since feature learning of

³<https://conceptnet.io>

Modality	Task	Dataset	prior Knowledge	Released
Attributes + Images	ZSL	AwA	text attributes for img/cls	2009
		AwA2	text attributes for img/cls	2019
		SUN	text attributes for img/cls	2012
		CUB	text attributes for img/cls	2010
	DG	Large-Scale Car	text attributes for img/cls	2017
Language + Images	Other	MS-COCO	text denotation graph	2014
		Flickr30K	text denotation graph	2015
		SBU Captions	text descriptions for img	2011
		Conceptual Captions	text descriptions for img	2018
Knowledge Graph + Images	ZSL	Visual Genome	flat concept graph	2017
		Mini-ImageNet	hierarchical concept graph	2016
		Tiered-ImageNet	hierarchical concept graph	2018
	DG	ImageNet	hierarchical concept graph	2009-2015
Images	ZSL	CIFAR-FS	N/A	2016
		FC-100	N/A	2016
	DG	Office-31	N/A	2010
		Office-Home	N/A	2016
		VisDA2017	N/A	2017

Table 3.2: Datasets for visual transfer learning and knowledge-based ML are summarized due to type of knowledge, task, prior knowledge, and their release date. We refer to tasks such as *zero-shot learning* (ZSL), *domain generalization* (DG), and *image classification*, *object detection*, *object segmentation*, and *image captioning* (Other).

images is an important part of DL and modern CV, this dataset is of limited use for end-to-end learned visual models.

AwA2 [157]: *Animals with Attributes 2* is a dataset for benchmarking transfer-learning algorithms, in particular for attribute-based classification and zero-shot learning. It served as a replacement for the original *Animals with Attributes* (AwA) dataset, with available raw images with public licenses from Flickr. The dataset has the same class structure and almost the same features. It consists of 37322 raw images of 50 animal classes. In alignment with AwA, the dataset provides pre-extracted feature representations for each image. AwA2 also provides a category-attribute matrix that contains an 85-part attribute vector for each category, e.g., color, stripes, fur, size, and habitat. Using the shared attributes, it is possible to transfer information between different classes.

CUB-200-2011 [158]: The CUB-200 [159] dataset and its extension the *Caltech-UCSD-Birds 200-2011* (CUB-200-2011) [158] dataset, are datasets describing 200 bird species. CUB-200-2011 has roughly double the number of images per class than CUB-200 and new part location annotations. It is fine-grained and medium-scaled concerning both the number of images and the number of classes, i.e. 11,788 images from 200 different categories, e.g., types of birds, annotated with 15 part locations, 312 attributes, and one bounding box per image. Akata et al. [160] introduces the first zero-shot

split of CUB with 150 training, 50 validation, and 50 test classes. It is also explicitly stated that the images in the dataset overlap with the images in ImageNet. Therefore, the results of transfer learning must be carefully checked when using networks previously trained with ImageNet or with images from Flickr, since the test set of CUB may overlap with the training set of the original network.

SUN Attribute Database [161]: SUN is a large-scale scene attribute database for high-level scene understanding and fine-grained scene recognition. It is crowdsourced by humans to find a taxonomy of 102 discriminative attributes. Next, the SUN attribute database was created on top of the fine-grained SUN categorical database. The attribute database consists of 14,340 images coming from 717 classes of scenes annotated with 102 attributes. The dataset is used to train attribute classifiers and evaluate how well these relatively simple classifiers can recognize a variety of attributes related to materials, surface properties, lighting, functions and affordances, and spatial envelope properties. Lampert et al. [162] use 645 classes of SUN for training, 65 classes for validation, and 72 classes for testing.

Fine-grained Car Dataset [163]: The *Fine-grained Car Dataset* originally consists of 2,657 classes and 700 thousand images. They refer to images from craigslist.com, cars.com, and edmunds.com as web images and those from Google Street View as GSV images. In addition to the category labels, each class is accompanied by metadata such as the make, model body type, and manufacturing country of the car. The dataset was adapted to domain generalization using a subset of 170 classes and 71,030 images [139]. The image category web images is used as the source domain, whereas the category GSV images suits as the target domain. The cars in web images are large and typically unoccluded, whereas those in GSV are small, blurry, and occluded.

Language-Augmented Image Datasets: We define datasets that provide context through textual descriptions in addition to the image as language-augmented image datasets. With the rise of multi-modal foundation models these datasets have increased their visibility recently. Besides the use of raw text descriptions, datasets have evolved that represent images and texts in a unified denotation graph.

MS-COCO [164]: *MS-COCO* includes images of complex everyday scenes with common objects in their natural context. It contains a total of 2.5 million labeled instances of 91 object types, in 328 thousand images. In addition, each image is described by five different human-written captions, e.g., "two men on motorcycles next to a stop sign.", "a motorcycle driver wearing a strange helmet at stop sign", or "a rider on a chopper with a helmet in the shape of a black skull with spikes". MS-COCO is used for class detection, instance detection, and instance segmentation. Recently, Zhang et al. [165] released an additionally learned denotation graph for MS-COCO, which structures text and

images into a combined representation. In addition, *MS-COCO* can be used for zero-shot learning tasks using the provided splits of unseen and seen class categories [166].

Flickr30K [167]: The *Flickr30K* is a standard benchmark for sentence-based image description and was originally designed for the tasks of image-based and text-based retrieval. The dataset contains 31 thousand images collected from the Flickr website, with five human-annotated text descriptions per image, and 276 thousand manually annotated bounding boxes. Each image is described independently by five annotators who are not familiar with the specific entities and circumstances, resulting in high-level descriptions such as “Three people setting up a tent”. All images are under the Creative Commons license. Moreover, they released a denotation graph for the dataset [165].

SBU Captions [168]: *SBU Captions* was originally designed for the purpose of generating text descriptions for images. It contains a large number of images from the Flickr website, where the images contain user-associated captions. The dataset was automatically crawled and filtered to produce a data collection containing over one million images with descriptive text descriptions. These text descriptions generally work similarly to captions and usually relate directly to some aspect of the visual image content, such as “Street dog in Lijiang”, “Fresh fruit and vegetables at the market in Port Louis Mauritius”, or “The sun was coming through the trees while I was sitting in my chair by the river”.

Conceptual Captions [169]: *Conceptual Captions* is a dataset for automatic image captioning provided by Google. It consists of about 3.3 million images equipped with text descriptions. In contrast to the curated style of the MS-COCO images, Conceptual Caption images and their raw descriptions are collected from the web. Therefore, it is an order of magnitude larger than MS-COCO and represents a wider variety of both images and image caption styles. More precisely, the raw text descriptions are taken from the Alt-Text of the HTML attribute associated with web images. In general, they automatically extracted, filtered, and transformed image and text descriptions, to achieve a balance between cleanliness, informativeness, fluidity, and learnability of the resulting captions ⁴. Additionally, Ng et al. [170] provided machine-generated labels for a subset of two million images from the Conceptual Captions training set.

Knowledge Graph-Augmented Image Datasets: We define KG-augmented image datasets as datasets that combine images with prior knowledge encoded in a KG. This prior knowledge in the KG can describe object and class relationships, higher-level concepts, or taxonomies. Therefore, every image of the dataset is connected to a graph representation that provides additional context for the understanding of the scene.

⁴<https://ai.google.com/research/ConceptualCaptions/>, accessed on 27 January 2023

Visual Genome [171]: *Visual Genome* provides a flat concept graph model of object relationships. These relationships describe the information that occurs in the image or even go beyond the information that can be extracted from the image. Visual Genome contains images with dense annotations of objects, attributes, and relationships modeled in a graph. More precisely, the dataset consists of over 100 thousand images, with each image containing an average of 21 objects, 18 attributes, and 18 pairwise relationships between objects. It summarizes objects, attributes, relationships, and noun phrases in region descriptions and map question answer pairs to WordNet synsets. Every image can contain multiple region graphs condensed of objects, attributes, and relationships, e.g., "stop sign is octagonal", "stop sign on the side of road", or "An old-fashioned car stopped in the street". All region graphs combined form a scene graph that describes all objects and their relations in the image. For zero-shot learning a split with 608 categories is considered for classification [166, 172]. Among these, 478 are seen categories, and 130 are unseen categories. This results in 54,913 training images and 7,788 test images. The relationship graph has 6,396 edges.

ImageNet [173]: The *ImageNet Large-Scale Visual Recognition Dataset and Challenge* is a benchmark in object category classification and detection on hundreds of categories and millions of images. The challenge has been run annually from 2010 to 2015. It contains one thousand classes and more than 1.2 million train, and 100 thousand test images per class for object classification. For the task of object detection, it contains one thousand classes and more than 450 thousand training images with 470 thousand bounding boxes, 50 thousand validation images with 55 thousand bounding boxes, and 40 thousand test images per class.

There are several derivatives of ImageNet with different appearances, such as *ImageNetV2* [174], *ImageNet Sketch* [175], *ImageNet-Vid* [176], *ImageNet Adversarial* [177], *ImageNet Rendition* [178], and such with synthetic distribution shifts, as *ImageNet-C* [119], and *Stylized ImageNet* [179]. More recently, a domain generalization scenario has been created in which ImageNet-trained models are tested on various ImageNet derivatives to evaluate the robustness of the models to distribution shift.

Mini-ImageNet [180]: *Mini-ImageNet* is a derivative of the ImageNet dataset and consists of 60 thousand color images of size 84×84 with 100 classes, each having 600 examples. Since this dataset fits in memory on modern computers, it is very convenient for rapid prototyping and experimentation. These 100 classes are divided into 64 train, 16 val, and 20 test classes for the zero-shot learning task.

Tiered-ImageNet [181]: *Tiered-ImageNet* is a subset of the ImageNet dataset. It groups classes into broader categories corresponding to higher-level nodes in the ImageNet hierarchy. There are 34 categories in total, with each category containing between 10 and 30 classes. For zero-shot learning, they split the categories into 20 training, six validation, and eight testing categories. This

ensures that all of the training classes are sufficiently distinct from the testing classes, unlike Mini-ImageNet.

Image Datasets without prior Knowledge

This section introduces transfer learning image datasets that have been originally created without prior knowledge. We list them because adding prior knowledge from generic KGs can be beneficial.

Zero-Shot Learning Datasets without prior Knowledge: We introduce image datasets without prior knowledge that have been applied mainly for zero-shot learning or few-shot learning tasks. As introduced in Section 3.1.1, zero-shot learning is part of an output domain change scenario.

CIFAR-FS [182]: *CIFAR-FS* is a few-shot learning dataset, randomly sampled from CIFAR-100 [183]. CIFAR-100, as well as CIFAR-10 [183], are labeled subsets of the *80 Million Tiny Images Dataset* [184], which was created in 2006 and withdrawn in 2020 due to inappropriate content. It contains 53,464 different nouns copied directly from Wordnet. Those terms were then used to automatically download images of the corresponding noun from Internet search engines and filters at the time to collect the 80 million images. *CIFAR-FS* contains 600 images for each of the 100 classes, which are further grouped into 20 superclasses. The limited original resolution of 32×32 makes the dataset well-suited for fast prototyping.

FC100 [185]: *Fewshot-CIFAR100* is also a derivative of the CIFAR-100 dataset and provides another few-shot learning split of the full CIFAR-100 dataset. The dataset is split into superclasses, rather than into individual classes to minimize the information overlap. Thus the train split contains 60 classes belonging to 12 superclasses, the validation and test contain 20 classes belonging to five superclasses each.

Domain Generalization Datasets without prior Knowledge: We provide a summary of image datasets without prior knowledge that have been applied mainly for domain generalization or domain adaptation tasks. As introduced in Section 3.1.1, domain generalization is part of an input domain change scenario.

Office-31 [186]: *Office-31* is an object recognition dataset that contains 31 categories and three domains, that is, *Amazon* (A), *Webcam* (W), and *DSLR* (D). These three domains have 2817, 498, and 795 instances, respectively. The images in Amazon are product images taken from amazon.com, the images in Webcam are the low-resolution images taken by web cameras, and the images in DSLR are the high-resolution images taken by DSLR cameras. In the experiments, every two of the three domains are selected as the source and the target domains, which results in six tasks. The evaluation contains all six cross-domain tasks: $A \rightarrow D$, $A \rightarrow W$, $D \rightarrow A$, $D \rightarrow W$, $W \rightarrow A$, $W \rightarrow D$.

Office-Home [187]: *Office Home* contains 15,585 images of 65 categories, collected from four domains: a) Art: 2421 artistic depictions of objects in the form of sketches, paintings, ornamentation, etc.; b) Clipart: a collection of 4379 clipart images; c) Product: 4428 images of objects without a background, akin to the Amazon category in the Office dataset; d) Real-World: 4357 images of objects captured with a regular camera. The evaluation contains all 12 cross-domain tasks.

VisDA2017 [188]: *The 2017 Visual Domain Adaptation Dataset and Challenge* is focused on the simulation-to-reality shift and has two associated tasks: image classification and image segmentation. The goal in both tracks is to first train a model on simulated, synthetic data in the source domain and then adapt it to perform well on real image data in the unlabeled test domain. VisDA2017 is the largest dataset for cross-domain object classification, with over 280 thousand images across 12 categories in the combined training, validation, and testing domains. The image segmentation dataset is also large-scale with over 30 thousand images across 18 categories in the three domains.

3.5 Summary

In this chapter, we introduced the task of transfer learning using prior knowledge. Therefore, we described the fundamentals of transfer learning and outlined different transfer learning scenarios based on the type of domain change. We categorized domain generalization and domain adaptation as tasks with an input domain change and zero-shot learning and few-shot learning as tasks with an output domain change. We provided insights into why DL suffers from domain-changing scenarios and introduced the specialization-generalization trade-off. In addition, we described how scale and inductive biases are used to enable DL methods to transfer. With interpreting inductive biases as infusion techniques for human knowledge we relate it to the idea of transfer learning using prior knowledge. We introduced transfer learning using KG as a promising variant of transfer learning using prior knowledge, since KGs are well suited to represent any kind of human understandable prior knowledge in explicit and implicit form. Finally, we provided relevant resources for visual transfer learning using KG, including generic KGs and datasets for visual transfer learning using prior knowledge.

Chapter 4

Combinations of Knowledge Graphs and Deep Learning

In recent years, the combination of KG and DL has gained increasing popularity for CV tasks. Whereas DL can learn rich representations from unstructured image data, KGs are well-known for representing explicit prior knowledge. Approaches of KG-DL aim to leverage the structured, human-engineered knowledge of KGs to enhance data-driven DL on images. However, integrating these two modalities of knowledge is not straightforward. While KGs encode their knowledge in graph-based explicit representations, DL approaches extract their knowledge and store vector-based representations in an implicit form.

As shown in Section 3.3.2, KGs are able to transform their explicit graph-based knowledge into implicit vector-based format. Therefore, KGE methods creates new opportunities for the field of combining KG and DL, by converting the KG into KGE \mathbf{h}_s , which enables the usage of linear operations and a combination with DNNs.

A common method for incorporating implicit prior knowledge encoded of a KGE with implicit knowledge of a DNN is to use a training objective. The training objective combines the semantic embedding \mathbf{h}_s from the KGE with the visual embedding \mathbf{h}_v created by the DNN. In this chapter, we introduce three distinct types of joint embeddings: a) semantic-visual embedding $\mathbf{h}_{s,v}$, where semantic data is embedded using \mathbf{h}_v as an objective; b) visual-semantic embedding $\mathbf{h}_{v,s}$, where visual data is embedded using \mathbf{h}_s as an objective; and c) hybrid embedding \mathbf{h}_h , where both semantic and visual data are embedded using a combination of \mathbf{h}_s and \mathbf{h}_v as an objective. Therefore, we collect approaches from different fields and different tasks to get a complete overview of KG-DL combinations. We provide a comprehensive summary and categorization of these approaches with respect to their integration of prior knowledge from the KG into the DNN.

In this chapter, we address the following research question:

RQ2: What are possible ways of integrating the prior knowledge encoded in a KG into the DL pipeline?

Our main contribution based on **RQ2** is as follows:

- *Categorization of methods combining KG and DL for visual transfer learning.*

The work of this chapter is mainly based on the following publication:

- **Sebastian Monka**, Lavdim Halilaj, and Achim Rettinger. 2022. A survey on visual transfer learning using knowledge graphs. *Semantic Web* 13, 3 (2022), 477–510.

The chapter starts with Section 4.1, that provides an overview of the methodology we used to conduct the work in this chapter, such as the strategy for literature search and selection of relevant works for the overview. Furthermore, in Section 4.2, a categorization is proposed based on four different categories of how a KG can be combined with a DL pipeline:

- 1) *Knowledge Graph as a Reviewer* - where the KG is used for post-validation of a visual model;
- 2) *Knowledge Graph as a Trainee*, where the KG is embedded into $\mathbf{h}_{s,v}$ using \mathbf{h}_v as objective;
- 3) *Knowledge Graph as a Trainer*, where the KG with \mathbf{h}_s is used as an objective to embed images into $\mathbf{h}_{v,s}$; and
- 4) *Knowledge Graph as a Peer*, where the KG with \mathbf{h}_s is combined with \mathbf{h}_v to suit as an objective that embeds both the KG and images into \mathbf{h}_h .

Since KGE methods have only recently entered the field of visual transfer learning, we also list related methods forming \mathbf{h}_s based on other types of prior knowledge in categories 2), 3), and 4). Other types of prior knowledge are language descriptions or class attributes so that their semantic features extractor $f_s(\cdot)$ differs in the type of input, but not in its architecture. Section 4.3 provides related work and Section 4.4 concludes the paper with a final summary.

4.1 Methodology

Our objective is to provide a comprehensive overview of how KGs can be used in combination with DL to solve visual transfer learning tasks. To ensure the quality of the outcome, we followed the process proposed by Petersen et al. [189, 190] and conducted an initial search on five scholarly indexing services. We applied inclusion and exclusion criteria on our findings and extended them based on the snowballing approach [191].

4.1.1 Literature Search

For the primary search process, we used five scholarly indexing services: Google Scholar¹, IEEE Xplore², ACM Digital Library³, Scopus⁴, and DBLP⁵. To collect relevant literature, we define a search string using the following strategy:

- Extract major terms from research questions.
- Use synonyms and alternative terms.
- Combine using *OR* to include synonyms and alternative terms.
- Combine using *AND* to join the key terms.

As a result, the following major terms related to the concepts are derived: Knowledge Graph, Visual Transfer Learning, and connect them by a Boolean AND operation. Each term contains a set of keywords related to the respective concept, connected by a Boolean OR operation. Therefore, the initial search string was as follows: (**"Knowledge Graph" OR "Knowledge Graph Embedding" OR "Semantic Embedding"**) AND (**"Visual Transfer Learning" OR "Transfer Learning" OR "Zero-shot Learning" OR "Deep Learning" OR "Computer Vision"**)

4.1.2 Literature Selection and Quality Assessment

After the literature search, we included literature based on the following criteria:

- Studies using visual features.
- Studies using prior knowledge.

Further, we excluded literature based on the following criteria:

- News articles.
- Non-English studies.
- Non-public available studies.
- Duplicate studies.

We reduced the amount of 16,200 studies after applying the inclusion and exclusion criteria on title and abstract to 17 relevant surveys and 164 studies (1.12%) During full-text reading, it

¹<https://scholar.google.com>

²<https://ieeexplore.ieee.org>

³<https://dl.acm.org>

⁴<https://www.scopus.com>

⁵<https://dblp.uni-trier.de>

became obvious that further articles should be removed as they were not in the scope based on the inclusion and exclusion criteria. The remaining articles (106) were used to conduct backward snowball sampling [191], which led to 22 additional studies. On the set of 128 primary studies, we conducted a quality assessment on the following questions:

- Does the study provide a solid assessment?
- Are the results plausible?

Thus, we were able to reduce the number of studies to 124. These studies provide the basis for the chapter and serve to answer the formulated research questions.

4.2 A Categorization of Visual Transfer Learning using KG

Visual transfer learning using KG has proven to be particularly advantageous compared to approaches without prior knowledge [96, 134]. Since prior knowledge mitigates the sole dependence on data distribution, it leads to models that are better generalized and thus more robust and applicable to new domains [128]. Having various kinds of prior knowledge, a KG can serve as a universal knowledge representation. KGs encode the classes either hierarchically, organized in superclasses, or flat, using relationships to other objects or other classes. Section 2.2.3 presents three distinct modeling structures with different levels of expressiveness and Section 2.3.2 introduces relevant embedding methods. All approaches that use a KG in combination with a DNN use the KG to implement some prior assumptions in the data-driven DL pipeline. A prior assumption induced by the KG is the definition of relationships between objects/classes so that objects/classes can borrow statistical strength from other related objects/classes in the graph. These priors give the CV process a structure that allows making better predictions even when visual data is sparse or erroneous. However, there are several ways the prior knowledge of a KG can be induced into a DNN.

This section provides a categorization of visual transfer learning approaches that combine KGs with the DL pipeline. As shown in Figure 4.1, we categorize the field of visual transfer learning using KG into:

- 1) *Knowledge Graph as a Reviewer* - where the KG is used for post-validation of a visual model;
- 2) *Knowledge Graph as a Trainee*, where a semantic-visual embedding $\mathbf{h}_{s,v}$ is learned using a visual embedding \mathbf{h}_v as objective;
- 3) *Knowledge Graph as a Trainer*, a visual-semantic embedding $\mathbf{h}_{v,s}$ is learned using a semantic embedding \mathbf{h}_s as objective; and
- 4) *Knowledge Graph as a Peer*, where a hybrid-embedding \mathbf{h}_h is learned using a combination of semantic embedding \mathbf{h}_s and a visual embedding \mathbf{h}_v as objective.

Since KGE methods have only recently entered the field of visual transfer learning, we also list related methods forming \mathbf{h}_s based on other types of prior knowledge in categories 2), 3), and 4). Other

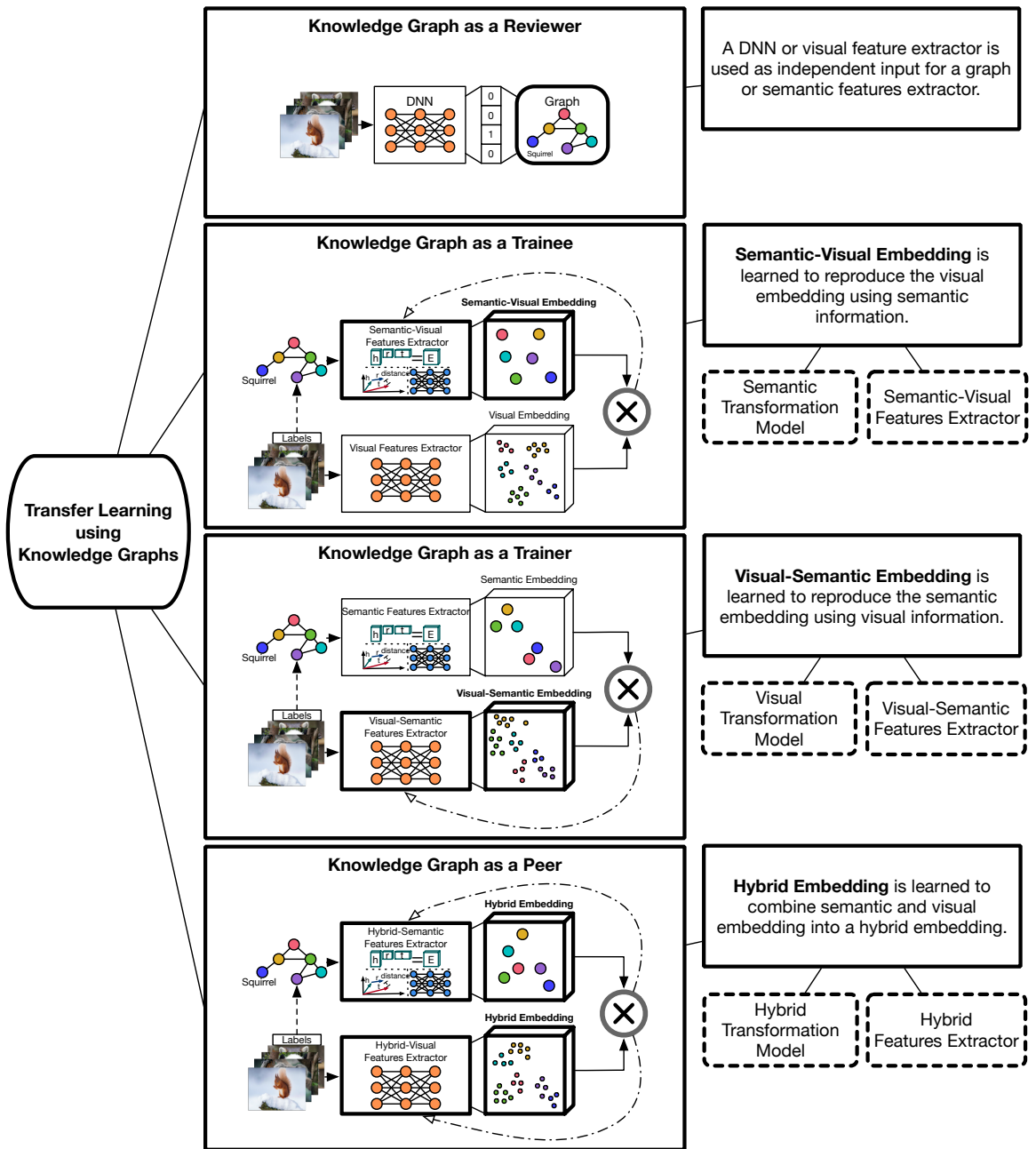


Figure 4.1: Visual transfer learning using KG is divided into four categories based on the role of the KG: 1) *Knowledge Graph as a Reviewer*; 2) *Knowledge Graph as a Trainee*; 3) *Knowledge Graph as a Trainer*; and 4) *Knowledge Graph as a Peer*. Each of the categories can be further separated into transformation models or feature extractors.

Category	Sub-Category	Output Domain Change	Input Domain Change	Other
Knowledge Graph as a Reviewer		[154], [192], [131], [132], [172]	[139], [193]	[194], [138], [140], [195], [196], [136], [143]
Knowledge Graph as a Trainee	Semantic-Visual Transformation Model	[197], [198]		
	Semantic-Visual Features Extractor	[134], [199], [137], [141], [200], [135], [201], [202], [203]		[204]
Knowledge Graph as a Trainer	Visual-Semantic Transformation Model	[160], [205], [96], [99], [206], [207], [208]		[209]
	Visual-Semantic Features Extractor	[210]	[211], [86]	[212]
Knowledge Graph as a Peer	Hybrid Transformation Model	[142], [213], [214], [215], [216], [217], [218], [219]	[213]	[220], [221], [222], [223]
	Hybrid Features Extractor	[224]		[225]

Table 4.1: Categories of KG-DL combinations and their tasks. Output domain change refers to category zero and few-shot learning, input domain change refers to the category domain generalization and adaptation, and other relates to object classification, object detection, and object segmentation on source task and domain only. Note: All approaches using related representations of prior knowledge instead of knowledge graphs are highlighted in red.

types of prior knowledge are language descriptions or class attributes, so that their semantic features extractor $f_s(\cdot)$ differs in the type of input, but not in its architecture, as described in Section 2.3.2.

We describe the categories and their approaches in detail and discuss their field of application and their properties. A summary of all approaches and their respective transfer learning task is given in Table 4.1.

4.2.1 Knowledge Graph as a Reviewer

Approaches of the category *Knowledge Graph as a Reviewer* arrange the visual model and the KG in sequential order, as depicted in Figure 4.2.

The visual output of a pre-trained DNN or its intermediate feature layers suit as an input to a graph or graph-based network. Unlike the other categories, the *KG as a Reviewer* does not learn a joint embedding space, instead, it uses the KG or its \mathbf{h}_s to reason over the independent output of a visual model \mathbf{h}_v .

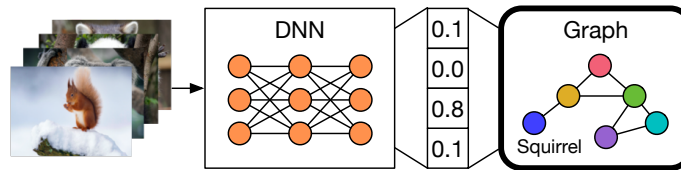


Figure 4.2: Approaches from the category *Knowledge Graph as a Reviewer* use the KG for post-validation of a pre-trained DNN or its intermediate feature layers.

Most of the approaches map the output of a visual features extractor $f_v(\cdot)$ on the corresponding input nodes in a hierarchical graph, to enrich the output with inter-class relationships. Lampert et al. [154] train a *support vector machine* (SVM) on SIFT features to predict binary AWA dataset attributes. These class attributes are fed into a hierarchical graph-based network to predict unknown classes for a zero-shot learning task. Salakhutdinov et al. [194] introduce a hierarchical Bayesian classification model [226] that learns a tree structure of class and super-class relationships. They use their learned graph on top of an SVM, which classifies HOG features of images. They show that their method using a learned graph outperforms a method using a fixed graph based on WordNet⁶ [148] and other approaches without hierarchical graph information. Deng et al. [192] proposed the *DARTS* algorithm for zero-shot learning. They pre-train an SVM on SIFT features of the ImageNet [130] dataset and map its classification output to WordNet with a reward and accuracy to maximize the information gain. Ordonez et al. [138] extend the approach to output human-understandable entry categories for images. They enrich the output of an SVM-based image classification model with information from a text-based n-gram language model by mapping both sources to the corresponding node in the WordNet graph. Rohrbach et al. [131] present *propagated semantic transfer* (PST). They use WordNet and attribute vectors from the AWA dataset to perform classification on few-shot learning classes of ImageNet. PST exploits similarities in visual embeddings of known classes encoded by an SVM learning a *k-Nearest Neighbor* (kNN) graph that helps to find relationships to new classes. Deng et al. [132] propose to use a *hierarchy and exclusion* (HEX) graph that exploits hierarchical class relationships of the output of a visual model. HEX graphs allow flexible specification of relations between labels applied to the same object. To build the graph, they use the hierarchical structure of WordNet extended with additional specifications and relations to objects, such as mutual exclusion (e.g., an object cannot be a dog and a cat), overlap (e.g., a husky can be a puppy and vice versa), and subsumption (e.g., all huskies are dogs). In addition, they proposed a probabilistic classification model that exploits their HEX graphs and evaluated their approach on ImageNet, in object classification and zero-shot learning. Gebru et al. [139] use WordNet attributes to improve fine-grained object classification on the task of domain generalization with the Office-31 [186] and the large-scale Car dataset [163]. Source and target domain images are fed through a

⁶<https://wordnet.princeton.edu/>

pipeline with two identical CNNs and a classification layer that classifies both the fine-grained classes and the different attribute types. The Kullback–Leibler divergence is used to compare the predicted label distributions. Lee et al. [133] propose a *graph gated neural network* (GGNN) that incorporates a structured KG based on WordNet and learned edge weights to improve zero-shot learning. First, an NN is learned that combines the GloVe [227] language embeddings of the class labels and the pre-trained visual feature vectors of the images as input to the GGNN. Second, the GGNN learns to propagate the information through the KG and outputs a probability for each node.

Instead of using hierarchical graphs of WordNet and class attributes only, other approaches make use of flat object or class relationships. Their graph consists of specific real-world configurations of objects and their appearance. Marino et al. [140] improves fine-grained image classification by creating a KG using the most common object-attribute and object-object relationships of the Visual Genome [171] dataset and higher-level semantics from WordNet. The output of a pre-trained, faster R-CNN [228] object detector is fed into a *graph search neural network* (GSNN) which reasons about relationships of the detected objects. The final prediction is a combination of the GSNN output, the visual embedding, and the detections of the faster R-CNN. Chen et al. [195] propose an object detection post-processing that connects a local and a global module via an attention mechanism. The local module is based on a convolutional *gated recurrent unit* (GRU) and builds spatial memory of previously detected objects using the class label and its visual embedding. The global graph-reasoning module consists of two paths, a spatial path that uses a region graph to connect far detected classes, and a semantic path that uses a KG, based on ADE20K [229] and Visual Genome, to connect classes with semantically related classes. Jiang et al. [196] extend [195] with *hybrid knowledge routed modules* (HKRM) allowing them to be applied on the intermediate feature representation directly to check the compatibility of prior knowledge with visual evidence in each image. HKRM can be divided into an explicit knowledge module and an implicit knowledge module, whereas the former contains external knowledge such as shared attributes, co-occurrence, and relationships, and the latter is built without explicit definitions and forms a region-to-region graph with constraints over objects, as spatial knowledge such as layout, size, overlap. Liu et al. [136] improve object detection by feeding the final object detections into a GCN which is based on object relationships and learned from MSCOCO dataset [164]. Gong et al. [193] propose a human parsing agent called "Graphonomy" that learns a KG on a conventional parsing network. It consists of an intra-graph reasoning module in form of a GCN whose structure uses semantic constraints from the human body to transfer knowledge within a dataset due to encoded relationships between nodes, and an inter-graph reasoning module, that uses handcrafted relations, a learnable matrix, feature similarities, and semantic similarities, to transfer semantic information between different datasets. Liang et al. [143] present a *symbolic graph reasoning* (SGR) layer for semantic segmentation and image classification. It consists of a module that assigns the visual features of a pre-trained DNN to corresponding nodes of a KG. Graph reasoning over all previously defined nodes is performed, and a

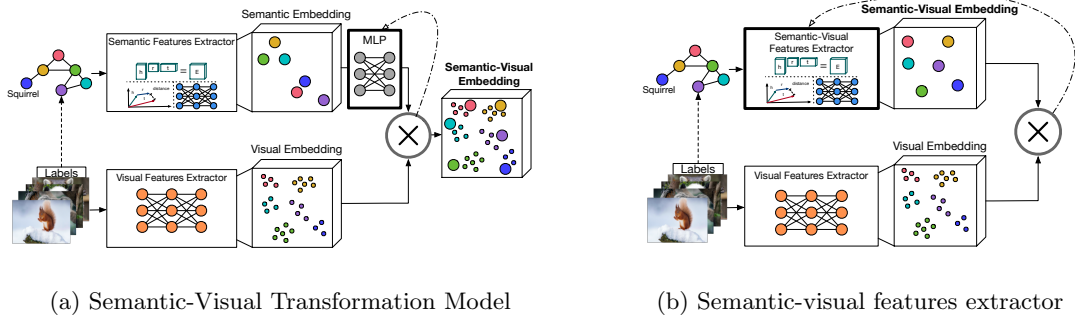


Figure 4.3: Approaches that belong to the category *Knowledge Graph as a Trainee* learn semantic visual embedding space supervised by a visual embedding. They either learn a) a transformation function, e.g. MLP, on top of a pre-trained semantic embedding space or b) a semantic-visual features extractor.

mapping from the symbolic graph information back to the visual feature space. Their graph is based on an object relation graph from Visual Genome and a hierarchical relation graph from WordNet. Luo et al. [172] propose a context-aware zero-shot learning framework, where they use a KG to reason about visual feature vectors generated from an object detection model. By using inter-class relationships, they improve traditional zero-shot learning techniques on the Visual Genome dataset.

4.2.2 Knowledge Graph as a Trainee

Approaches that belong to this category combine the visual DNN with the prior knowledge of a KG by learning a semantic-visual embedding $\mathbf{h}_{s,v}$. Unlike the *Knowledge Graph as a Reviewer*, which uses the visual embedding \mathbf{h}_v as input for the KG, approaches from the category *Knowledge Graph as a Trainee* use \mathbf{h}_v as an objective to embed the KG into $\mathbf{h}_{s,v}$.

Figure 4.3 illustrates a conceptual architecture of the *Knowledge Graph as a Trainee* approach. To combine visual and semantic information, some approaches either learn a transformation function, e.g. MLP, on top of a semantic embedding space \mathbf{h}_s , or apply supervised KGE methods to learn a semantic-visual features extractor $f_{s,v}(\cdot)$ directly.

Semantic-Visual Transformation Models

As shown in Figure 4.3a, the pre-trained \mathbf{h}_s is fixed over the whole training process, and an additional transformation function, e.g. MLP, is learned to transform \mathbf{h}_s into the semantic-visual embedding space $\mathbf{h}_{s,v}$.

Related Approaches using other Prior Knowledge: Roohan et al. [197] used a fixed language embedding to define relationships between classes, that unknown classes in a zero-shot learning task can borrow their visual embeddings from a linear combination of known related classes. Zhang et

al. [198] extends the method by suggesting to use the visual space, instead of the semantic space, as the main embedding space. They claim that this will reduce the hubness problem that occurs in high dimensions.

Semantic-visual Features Extractors

As illustrated in Figure 4.3b the semantic-visual features extractor $f_{s,v}(\cdot)$ learns to directly transform the KG into a semantic-visual embedding $\mathbf{h}_{s,v}$ using the supervision of the visual embedding space \mathbf{h}_v . As described in Section 2.3.2, $f_{s,v}(\cdot)$ is mostly implemented using a supervised KGE method.

Wang et al [134] build a GCN on the structure of WordNet and optimize it to predict ImageNet pre-trained visual classifiers. Based on the learned relations in the GCN they are able to transform information into novel class nodes to perform zero-shot learning. A similar principle is used by Chen et al. [204] for multi-label image recognition. However, instead of using a hierarchical graph, the approach uses an object-relation graph which reflects the different relations between objects in a scene. They build their graph based on the occurrence probabilities of different objects in the MSCOCO dataset since some objects are more likely to occur together. Kampffmeyer et al. [199] claim that multi-layer GNN architectures, which are required to propagate knowledge to distant nodes in the graph, dilute the knowledge by performing extensive Laplacian smoothing at each layer and thereby consequently decrease performance. They propose a *dense graph propagation* (DGP) module with direct links among distant nodes to exploit the hierarchical graph structure of the KG. They tested their approach on zero-shot learning tasks such as the 21K-ImageNet dataset and AWA2. Gao et al. [137] designed a *two-stream GCN* (TS-GCN) to perform *zero-shot action recognition* (ZSAR). Their GCN architectures are based on the ConceptNet 5.5 KG, which contains information from various knowledge bases such as WordNet and DBpedia. The first classifier branch uses the language embedding vectors of all classes as input for a GCN and then generates the classifiers for each action category. The second instance branch feeds video segments into a DNN and outputs object scores, which are combined with attribute vectors from the classifier branch using a post-processing GCN to form an attribute feature space. The final objective is then defined by a comparison of the attribute feature space and the output of the classifier branch. Peng et al. [141] propose a *knowledge transfer network* (KTN), which extends [134] with a vision-knowledge fusion model. This vision-knowledge fusion model is used to combine the final prediction output of the GCN with the output of a DNN, as they claim that semantic embeddings and visual embeddings are complementary and therefore cannot be combined with a single inner product. They pre-train their visual feature learning module using cosine similarity on image data, use a subgraph of WordNet for their knowledge transfer module, and language embeddings of the class labels as the initial state of the nodes of the GCN. Chen et al. [200] present the *knowledge graph transfer network* (KGTN). The knowledge graph transfer module incorporates a GGNN, which supports a knowledge transfer of classes through a KG. To train GGNN, they fix the weights of a pre-trained visual features extractor

and examine three different similarity metrics, such as inner product, cosine similarity, and person correlation coefficient, to compare the output of the DNN and the GGNN. They show that the accuracy of the model benefits from a reasoning process and the prior knowledge from a KG.

Geng et al. [135] recently proposed Onto-ZSL, an ontology-enhanced zero-shot learning framework that can be applied either to image classification or knowledge graph completion. They build an inter-class relationship using an ontological schema, that comprises a label taxonomy from WordNet, as well as text and attribute descriptions. Further, they address the data imbalance problem between seen and unseen images by leveraging a *generative adversarial network* (GAN) that produces synthesized visual feature vectors for unseen classes.

Related Approaches using other Prior Knowledge: Approaches using language models leverage GANs to imagine unseen categories from text descriptions and hence recognize novel classes with no examples being seen. GANs can be seen as a transformation function from text-based input to visual features, using the supervision of a visual model. Zhu et al. [201] propose GAZL, an approach that takes noisy text descriptions about unseen classes from Wikipedia and generates synthesized visual features for this class. Using textual input for unseen classes they learn a GAN that generates visual features similar to the pre-trained ones of the seen classes. Therefore, the zero-shot learning problem is transformed into a standard classification task and a classifier that can handle unseen classes can be trained using the synthesized image features for every unseen class. Li et al. [202] extended the approach by introducing LisGAN, a GAN that takes semantic descriptions and random noise to generate visual features for unseen classes. In addition, they deploy the average representation of all samples from an unseen class defining the sole sample of the class to reduce the noise in the predictions. Vyas et al. [203] propose LsrGAN, a generative model that leverages the semantic relationship between seen and unseen categories and explicitly performs knowledge transfer by incorporating a novel *semantic regularized loss* (SR-Loss). Knowing the inter-class relationships in the semantic space helps to impose the same relationship constraints to the generated visual features.

Methods that belong to the category *Knowledge Graph as a Trainer* combine the visual output of a DNN with the prior knowledge of a KG by learning a visual-semantic embedding $\mathbf{h}_{v,s}$.

4.2.3 Knowledge Graph as a Trainer

Figure 4.4 illustrates a conceptual architecture of the *Knowledge Graph as a Trainer* approach. The KG acts as a trainer and supervises the training of the DNN using \mathbf{h}_s , rather than letting the DNN learn a \mathbf{h}_v solely depending on the data distribution of the images. We refer to such an embedding of visual information learned under the supervision of a semantic embedding \mathbf{h}_s as a visual-semantic embedding $\mathbf{h}_{v,s}$. To combine semantic and visual information, some approaches either learn a transformation function, e.g. MLP, on a pre-trained and fixed visual embedding \mathbf{h}_v or learn a visual-semantic features extractor $f_{v,s}(\cdot)$ directly.

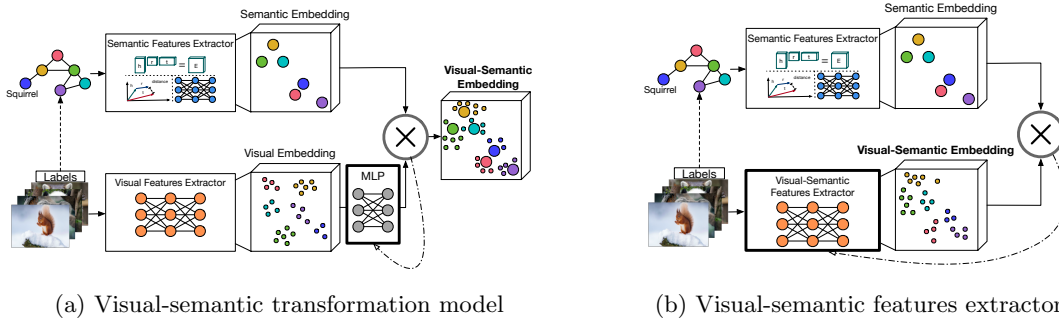


Figure 4.4: Approaches that belong to the category *Knowledge Graph as a Trainer* learn visual semantic embedding space supervised by a semantic embedding. They either learn a) a transformation function, e.g. MLP, on top of a pre-trained visual embedding space that suits as a transformation function or b) a visual-semantic features extractor that learns the final embedding directly.

Visual-Semantic Transformation Models

As shown in Figure 4.4a, the pre-trained \mathbf{h}_v is fixed over the whole training process and a transformation function, e.g. MLP, is learned to transform \mathbf{h}_v , into the visual-semantic embedding $\mathbf{h}_{v,s}$.

Akata et al. [160] refer to their semantic embedding space transformations as label embedding methods. They compared transformation functions from the visual embedding space to the attribute label embedding space, the hierarchy label embedding space, and the Word2Vec [230] label embedding space. Lonij et al. [144] approached the task of open-world visual recognition by using KGs. They learn \mathbf{h}_s from a WordNet KG by using the *neural tensor layer* (NTL) [231] architecture and embed the visual embedding generated by a pre-trained CNN into the same space using the hinge rank loss.

Related Approaches using other Prior Knowledge: One of the first approaches that use semantic embeddings with NNs is the work from Mitchell et al. [209]. They use language embeddings derived from text corpus statistics to generate neural activity pattern images. Instead of generating images from text, Palatucci et al. [205] learn a linear regression model to map neural activity patterns into language embedding space. Socher et al. [96] present a model for zero-shot learning that learns a transformation function between a visual embedding space, obtained by an unsupervised feature extraction method, and a semantic embedding space, based on a language model. The authors trained a 2-layer NN with the MSE loss to transform the visual embedding into the language embedding of eight classes. Frome et al. [99] introduce the deep visual-semantic embedding model DeViSE that extends the approach from eight known and two unknown classes to one thousand known and 20 thousand unknown classes. Therefore, they pre-train their visual features extractor using ImageNet and their semantic embedding vector using a skip-gram language model [230]. In contrast to Socher et al. [96] they learn a linear transformation function between

the visual embedding space and the semantic embedding space using a combination of dot-product similarity and hinge rank loss since they claim that MSE distance fails in high-dimensional space. Norouzi et al. [206] propose *convex combination of semantic embeddings* (ConSE). ConSE performs a convex combination of known classes in the semantic embedding space, weighted by their predicted output scores of the DNN, to predict unknown classes in a zero-shot learning task. Similarly, Zhang et al. [207] introduce the *semantic similarity embedding* (SSE), which models target data instances as a mixture of seen class proportions. They built a semantic space in that each novel class could be represented as a probabilistic mixture of the projected source attribute vectors of the known classes.

Kodirov et al. [208] propose SAE a semantic autoencoder for zero-shot learning. It is learned by encoding pre-trained visual features of a CNN into a latent semantic space and then by decoding them back into visual space. The semantic space is based on class attributes for smaller datasets and on a word2vec language model for larger datasets. They claim that their latent semantic embedding space can better handle the projection domain shift problem, i.e. the distribution shift between seen and unseen classes.

Visual-semantic Features Extractors

As illustrated in Figure 4.4b the visual-semantic features extractor $f_{v,s}(\cdot)$ is learned to directly transform the images into a visual-semantic embedding $\mathbf{h}_{v,s}$ using the supervision of the semantic embedding space \mathbf{h}_s . As described in Section 2.3.2, \mathbf{h}_s is mostly learned using an unsupervised KGE method and $f_{v,s}(\cdot)$ is implemented using a standard DNN.

Monka et al. [211] propose KG-NN, an approach that uses a KG and its \mathbf{h}_s to train a visual DNN. Using a contrastive knowledge graph embedding loss in combination with \mathbf{h}_s they learn a visual-semantic features extractor $f_{v,s}(\cdot)$. They test their approach on domain generalization and adaptation tasks for road sign recognition in Germany and China, as well as on Mini-ImageNet and various derivatives. They show that their visual features extractor learned using the *Knowledge Graph as a Trainer* outperforms a conventional DNN trained with CE, the same DNN without additional information from the KG, and the same DNN using additional information from a pre-trained GloVe embedding in visual transfer learning tasks.

Jayathilaka et al. [210] proposed a framework named ViOCE that integrates ontology-based background knowledge in the form of n-ball class embeddings into a DNN-based vision architecture. The approach consists of two components - converting symbolic knowledge of an ontology into continuous space by learning n-ball embeddings that capture properties of subsumption and disjointness and guiding the training and inference of a vision model using the learned embeddings.

Related Approaches using other Prior Knowledge: Joulin et al. [212] demonstrate that feature extractors trained to predict words in image captions learn useful image representations. They convert the title, description, and hashtag metadata of images into a bag-of-words multi-label

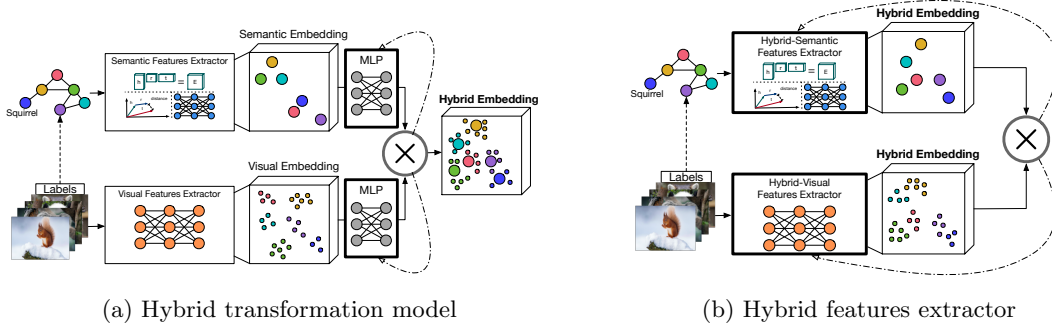


Figure 4.5: Approaches that belong to the category *Knowledge Graph as a Peer* learn hybrid embedding space as a combination of visual and semantic embedding space. They either learn a) transformation functions, e.g. MLPs, on top of both pre-trained visual and semantic embedding spaces that suit as a transformation function or b) hybrid features extractors that learn the final embedding directly.

classification task and showed that pre-training a feature extractor to predict these labels learned representations which performed similarly to ImageNet-based pre-training on transfer tasks. Radford et al. [86] claim that state-of-the-art CV systems are restricted to predict a fixed set of predetermined object categories. Therefore, they propose to use a simple and general pre-training of their CNN with natural language supervision, i.e. predicting which caption goes with which image on a dataset of 400 million image-text pairs collected from the Internet using the objective of Zhang et al. [225].

Approaches of the category *Knowledge Graph as a Peer* combine the visual DNN with the prior knowledge of a KG by influencing both semantic and visual embedding. Unlike the previous categories, the idea of a hybrid embedding \mathbf{h}_h is to fuse the visual embedding \mathbf{h}_v and the semantic embedding \mathbf{h}_s . Both semantic and visual data are then embedded into \mathbf{h}_h .

4.2.4 Knowledge Graph as a Peer

Figure 4.5 illustrates a conceptual architecture of the *Knowledge Graph as a Peer* approach. The final hybrid embedding space is either a combination of pre-trained visual embedding \mathbf{h}_v and semantic embedding \mathbf{h}_s , using a transformation function, e.g. MLP, or a combination of hybrid-visual $f_{h,v}(\cdot)$ and hybrid-semantic features extractors $f_{h,s}(\cdot)$.

Hybrid Transformation Models

As shown in Figure 4.5a, pre-trained \mathbf{h}_s and pre-trained \mathbf{h}_v are fixed over the whole training process and additional transformation functions, e.g. MLPs, are learned to transform \mathbf{h}_s and \mathbf{h}_v , into the hybrid embedding space \mathbf{h}_h .

Zhao et al. [142] propose a joint model that combines an image stream and a concept stream via a joint loss function to preserve concept hierarchy as well as visual feature similarities. The

concept stream is based on a language embedding with the hierarchical graph of WordNet and the image stream is a visual embedding from semantic segmentation DNN. They compare their approach against the standard CE-based approach and semantic embedding space transformations based on Word2Vec. Roy et al. [214] introduce a zero-shot learning model that takes advantage of the commonsense KG ConceptNet 5.5 to generate \mathbf{h}_s of the class labels by using a GCN-based autoencoder. They enrich \mathbf{h}_s with additional attributes and language embeddings, which is then compared with a pre-trained visual output of a DNN using a relation network [232].

Related Approaches using other Prior Knowledge: Yang et al. [213] propose a two-sided NN to learn a combination of a pre-trained visual embedding and a semantic embedding of attributes and word vectors based on image descriptions to perform zero-shot learning and domain generalization. To train their NN they use a Euclidean loss for regression and a hinge rank loss for classification. Fu et al. [215] try to reduce the bias of semantic embedding spaces, by proposing a transductive multi-view embedding framework that aligns novel features with the semantic embedding space for zero-shot learning. The framework first transforms the semantic embedding space into a joint embedding space using the unlabeled target data with a multi-view *canonical correlation analysis* (CCA) to alleviate the projection domain shift problem. And Second, a heterogeneous multi-view hypergraph label propagation method is used to perform zero-shot learning in the transductive embedding space, which combines additional semantic knowledge in the form of attributes and word vectors from related classes. Ba et al. [216] introduce a flexible zero-shot learning model that learns to predict unseen image classes using a language embedding. Therefore, they add two separate MLPs on top of the visual embedding and the semantic embedding and train them using the binary-CE loss, the hinge loss, and the Euclidean distance loss. Karpathy et al. [220] learn a model that generates language descriptions for detected objects in an image. Their objective aligns the output of a pre-trained CNN applied to image regions, and the output of a bidirectional RNN applied to sentences. Changpinyo et al. [217] use a set of “phantom” object classes whose coordinates live in both the semantic space and the model space. To align the two spaces, they view the coordinates in the visual embedding as the projection of the vertices on the graph from the semantic embedding. To compute low-dimensional Euclidean space embeddings from the weighted graph they propose to use the algorithm of Laplacian eigenmaps, mapping semantic and visual embedding into a common space defined by the mixture of seen classes proportions. Tsai et al. [218] propose the approach ReViSE that learns an unsupervised joint embedding of semantic and visual features to enable zero-shot learning. As external knowledge, they experiment with three different embedding methods for their attributes, human-annotated attributes [162], Word2Vec attributes, and GloVe attributes. Tang et al. [221] propose the *large scale detection through adaptation* (LSDA) framework to improve object detectors with image classification DNNs, hence without requiring expensive bounding box annotations. LSDA defines visual similarity as the distance between pre-trained visual embedding vectors and semantic similarity as the distance between pre-trained language embedding vectors

of the labels. Jiang et al. [219] introduce their *transferable contrastive network* (TCN) explicitly transfers knowledge from the source classes to the target classes, to counteract the overfitting problem on source classes. To compute the similarities between classes in the hybrid embedding space, they design a contrastive network that automatically judges how well the embedding vector is consistent with a specific class. Li et al. [222] propose a multi-layer transformer [53] model as DNN, which uses object tags detected in images as anchor points to learn a joint embedding of the detected objects and the language tags, instead of simply concatenating visual embedding and semantic embedding. Yu et al. [223] propose a knowledge-enhanced approach, ERNIE-ViL, to learn joint representations of vision and language using a transformer model as DNN. ERNIE-ViL tries to construct detailed semantic connections across vision and language while constructing a scene graph parsed from sentences and type prediction tasks, i.e., object prediction, attribute prediction, and relationship prediction in the pre-training phase.

Hybrid Features Extractors

As depicted in Figure 4.5b, hybrid-semantic $f_{h,s}(\cdot)$ and hybrid-visual $f_{h,v}(\cdot)$ features extractors are learned to directly transform KG and images into a common hybrid embedding \mathbf{h}_h . As described in Section 2.3.2, $f_{h,s}(\cdot)$ is usually implemented using a supervised KGE method and $f_{h,v}(\cdot)$ using a standard DNN.

Recently, Naeem et al. [233] proposed a method to perform zero-shot image classification using hybrid features extractors. An ImageNet pre-trained DNN is used for the visual features extractor and a GCN in the *compositional graph embedding* (CGE) setting is used for the semantic features extractor. However, they learn a joint embedding function that can influence the weights of the DNN as well as the weights from the GCN. Interestingly, they compare their model against a similar version of their model, but with a fixed visual features extractor where the KG just acts as a trainee, as proposed in Section 4.2.2. They use that version for comparison with related approaches, stating that all other methods are based on fixed visual features extractors. Moreover, they show that a hybrid approach with an adaptive visual features extractor performs better than the other.

Related Approaches using other Prior Knowledge: Zhang et al. [225] use two contrastive pre-training objectives, contrasting semantic embedding to visual embedding, and vice versa, on the special domain of medical imaging to learn a joint feature extractor. Instead of previous works that learn transformation functions on top of fixed image-trained visual features extractors they directly supervise the training of the CNNs with language embedding information. To train their DNN they use text-image paired data.

4.3 Surveys Related to Visual Transfer Learning using KG

This section discusses related work for this chapter. Therefore, we outline surveys from visual transfer learning and knowledge-based machine learning. Furthermore, we provide additional insight into surveys on the topic of explainable AI, as the field is strongly related to knowledge-based ML.

Visual Transfer Learning: Pan et al. [113] and Zhang et al. [234] categorized the task of visual transfer learning into three main settings: inductive, transductive, and unsupervised transfer learning. In inductive transfer learning the task changes from source to target, whereas the domain stays the same. In transductive transfer learning, the source and target tasks are the same, while the source and target domains are different. Finally, in the unsupervised transfer learning setting, similar to inductive transfer learning, the target task is different from but related to the source task. However, unsupervised transfer learning focuses on solving learning tasks when no labeled data is available in the source and the target domain. Weiss et al.[235] separated the field into homogeneous and heterogeneous transfer learning, whereas approaches of the former are developed and proposed for handling the situations where the domains are of the same feature space and the latter refers to the knowledge transfer process in the situations where the domains have different feature spaces. Kaboli et al. [236] reviewed and structured 20 transfer-learning approaches. Wang et al. [237] investigated the field from the domain change perspective. If the domain change is small they call it homogeneous transfer learning and if the domain change is large they call it heterogeneous transfer learning. Zhang et al. [238] further separated the field of transfer learning into 17 different tasks, based on supervision, the amount of labeled data, and the size of the domain gap. Zhang et al. [234] categorized transfer learning based on their adaptation process into weakly supervised learning, instance re-weighting, feature adaptation, classifier adaptation, deep network adaptation, and adversarial adaptation. Wang et al. [239] provides a comprehensive survey about zero-shot learning methods and their different semantic spaces. These semantic spaces can either be engineered semantic spaces, generated by attributes, lexicals, and text keywords, or learned semantic spaces, such as label-embeddings, text-embeddings, and image-representations. Xian et al. [157] recently released a survey about zero-shot learning which structures the field into methods that learn linear compatibility, nonlinear compatibility, intermediate attribute classifier, or hybrid models.

Knowledge-Based Machine Learning: Only a few surveys have investigated the field of knowledge-based ML. Von Rueden et al. [84] recently published a survey about knowledge-based ML under the term *informed machine learning*. They structure the field based on the source of the knowledge, the representation of the knowledge, and the integration of the knowledge into the ML pipeline. Further, Gouidis et al. [240] structured the knowledge-based ML literature into approaches with symbolic knowledge, commonsense knowledge, and the ability to learn new knowledge. They give an overview of different works that combines ML with knowledge-based approaches in the field of

CV. They categorized the approaches due to their CV task, e.g. object detection, scene understanding, image classification, their applied ML architecture, e.g. CNN, GNN, RCNN, and their loss function, e.g. scoring functions, probabilistic programming models, and Bayesian Networks. Ding et al. [241] reviewed all ontology applications in the field of object recognition.

Another research field in demand is *Explainable AI*, where knowledge-based methods and ML approaches are combined. Explainable AI refers to methods and techniques of ML such that the results of the solution can be understood by humans. Futia et al. [242] investigated the field of explainable AI using KGs and categorized approaches into knowledge matching, cross-disciplinary and interactive explanations. Chen et al. [243] and Chari et al. [244] proposed to use hybrid explanations of a taxonomy generated for the end-user, including causal methods, neuro-symbolic AI systems, and representation techniques. Seeliger et al. [245] summarized semantic web technologies that can provide valid explanations for ML models, separating them due to their ML technique and semantic expressiveness. Chen et al. [246] recently proposed a survey about knowledge-aware zero-shot learning. They divided the ML methods that approach the zero-shot learning task into three distinct categories: mapping function-based, generative model-based, and graph neural network based. In their work, they additionally provide an overview of different types of prior knowledge, e.g. text, attribute, KG, and rule and ontology.

Aditya et al. [247] provide a survey about reasoning mechanisms and knowledge integration methods for image understanding applications. Besides an overview of frameworks that handle logic operations, they briefly discuss at which position prior knowledge can be introduced into a DL pipeline: i) Ahead of the DNN, through a pre-processing of domain knowledge and augmentation of training samples; ii) Inside the DNN, through a vectorization of parts of the knowledge base and as an input to intermediate layers; iii) Inside of the DNN, to inspire the neural network architecture; and iv) After the DNN, as a post-processing using external knowledge.

We understand their taxonomy as a general explanation of where external knowledge can be induced into the DL pipeline. For instance, our category *Knowledge Graph as a Reviewer* is related to iv), since the KG can operate as a post-processing network on the output of the visual DNN. However, we also see that the reasoning process of the *Knowledge Graph as a Reviewer* can be applied on an intermediate visual feature layer of the DNN. Similarly, the categories *Knowledge Graph as a Trainee*, *Knowledge Graph as a Trainer*, and *Knowledge Graph as a Peer* overlap with categories ii) and iii).

However, in contrast to Aditya et al., our categories are described by the explicit information exchange between the visual and semantic embedding space. Instead of a categorization based on the position of the knowledge induction, our categories depend on whether the semantic embedding inspires the visual embedding or vice versa. Using our categories, we, therefore, describe four distinct principles used to combine the two modalities.

4.4 Summary

Visual transfer learning using prior knowledge has gained increasing attention in research. Since initiatives for building and maintaining generic KGs host a large research community, we believe that exploiting them with DL will improve various applications, especially in visual transfer learning. This chapter presented a deep analysis of existing approaches and investigated how KGs, as a unified representation of prior knowledge, can be utilized in various forms. The insights gained from this analysis can be valuable in developing solutions for addressing challenges and open issues in the field. The main contributions of the chapter are framed in **RQ2**, and we summarize the answer as follows:

- **What** are possible ways of integrating the prior knowledge encoded in a KG into the DL pipeline?

Approaches of the field of visual transfer learning using KG can be separated into four distinct categories based on how the KG is combined with the DL pipeline:

- 1) *Knowledge Graph as a Reviewer* - where the KG is used for post-validation of a visual model;
- 2) *Knowledge Graph as a Trainee*, where a semantic-visual embedding $\mathbf{h}_{s,v}$ is learned using a visual embedding \mathbf{h}_v as objective;
- 3) *Knowledge Graph as a Trainer*, a visual-semantic embedding $\mathbf{h}_{v,s}$ is learned using a semantic embedding \mathbf{h}_s as objective; and
- 4) *Knowledge Graph as a Peer*, where a hybrid-embedding \mathbf{h}_h is learned using a combination of semantic embedding \mathbf{h}_s and a visual embedding \mathbf{h}_v as objective.

- **What** are the properties of the respective combinations?

It can be seen that every category has its applications in distinct tasks.

- 1) *Knowledge Graph as a Reviewer* - approaches leverage prior knowledge by using it as an independent post-validation. The KG or KGE (\mathbf{h}_s) enables reasoning over the output or intermediate feature layers of the DNN. However, the modalities are either learned independently or in sequential order, such that semantic and visual embedding spaces are not directly influenced by each other.
- 2) *Knowledge Graph as a Trainee* - approaches leverage prior knowledge by providing a structure for a KGE method, e.g. GNN, that is learned using \mathbf{h}_v as objective. Approaches are used mainly in the zero-shot learning scenario to extend the learned model to classes that are not present in the training data, using the inductive property of GNNs combined with the ability of DNNs to extract relevant features of images.

3) *Knowledge Graph as a Trainer* - approaches leverage prior knowledge by influencing DNNs in learning specific visual features. The DNN can learn an image data distribution independent embedding provided by \mathbf{h}_s instead of just using the data distribution. Thus, we see the advantage of these approaches specifically in the domain generalization scenario.

4) *Knowledge Graph as a Peer* - approaches leverage prior knowledge by influencing semantic and visual embedding equally. Although it is not clear which modality dominates the other and therefore the learned embedding, approaches have yielded quite promising results for zero-shot learning and domain generalization tasks.

Chapter 5

Learning Visual Models using a Knowledge Graph as a Trainer

As shown in Chapter 3, transfer learning is a ubiquitous concept in DL. Since DL learns its model on a constrained training dataset, it must always deal with a trade-off between specialization and generalization. In particular, in transfer learning scenarios where the domain gap between source and target domains continues to increase, DL methods require additional knowledge to support generalization [110]. To reduce the high dependency on the training domain, pre-training methods that generate rich embedding spaces seem to be a promising research direction for CV and NLP. Exploring these embedding spaces, it is found that DNNs encode visually similar classes close to each other when sufficient training data is available. Recently, the idea of training a DNN with an image-independent embedding space in form of language embeddings has also been proven to be beneficial for CV tasks [212, 86, 225].

In this Chapter, we introduce the *knowledge graph neural network* (KG-NN), a novel approach to learn a visual model using a KG and its KGE \mathbf{h}_s as a trainer. More concretely, a domain-invariant embedding space using a KG and an appropriate KG embedding algorithm is constructed. We then train KG-NN with a contrastive loss function to adapt its visual embedding to \mathbf{h}_s given by the KG. KG-NN, therefore, learns the relevant features of the images by connecting semantically similar classes and distinguishing them from different ones. The benefit is two-fold. First, KG-NN will be more robust to distribution shifts since the embedding space is independent of the dataset distribution, and second, it is enriched with additional semantic data in a controlled manner.

To investigate the generalization and adaption of KG-NN in real-world scenarios, the task of visual transfer learning provides a suitable testing environment. Transfer learning tasks consist of a source and a target dataset, differing in terms of their underlying distribution, e.g sensors, environments, and countries. A domain generalization task has only access to labeled source data,

whereas the domain adaptation task contains a small amount of additional labeled target data. For domain generalization - *Scenario 1*, we performed two experiments: 1) object classification, where the DNN is trained on the Mini-ImageNet [180] dataset and evaluated on derivatives; 2) road sign recognition, where the DNN is trained on the *German Traffic Signs Dataset* (GTSRB) [248] and evaluate on the *Chinese Traffic Signs Dataset* (CTSD) [249]. For domain adaptation - *Scenario 2*, we train the DNN on GTSRB and additional labeled target data from CTSD. In both scenarios, the respective KGs are developed in RDF representation. RDF provides the necessary means for an easy and flexible extension of the defined schemas and allows for enriching and interlinking entities in the KGs with complementary information from other sources. The generality of our approach becomes apparent in the fact that it can be assigned to any method using DNNs since we provide an alternative and enriched training method. Our results indicate that KG-NN is significantly more accurate compared to a conventional approach based on the cross-entropy loss in any domain-changing scenario.

This chapter introduces a novel approach called KG-NN, which combines KGs with DL architectures to enhance visual transfer learning. Hereby, KG-NN is learned to align the visual embedding space \mathbf{h}_v to the semantic embedding space \mathbf{h}_s of the KGE.

In this chapter, we address the following research question:

RQ3: How can prior knowledge encoded in a KG guide DL to improve visual transfer learning?

Our main contribution of this paper is summarized as follows:

- *Method to learn a DNN using a KG as a Trainer.*

This contribution can be divided into the following sub-contributions:

- We introduce KG-NN, a neuro-symbolic approach that uses prior domain-invariant knowledge captured by a KG to train a DNN.
- We adapt a contrastive loss function to combine knowledge graph embeddings with the visual embeddings learned by the DNN.
- We evaluate the KG-NN approach in domain generalization and domain adaptation tasks on two different scenarios with respective image datasets.

The work in the chapter is mainly based on the following publication:

- **Sebastian Monka**, Lavdim Halilaj, Stefan Schmid, and Achim Rettinger. 2021. Learning Visual Models Using a Knowledge Graph as a Trainer. In *The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24-28, 2021, Proceedings (Lecture Notes in Computer Science)*, Springer, 357–373.

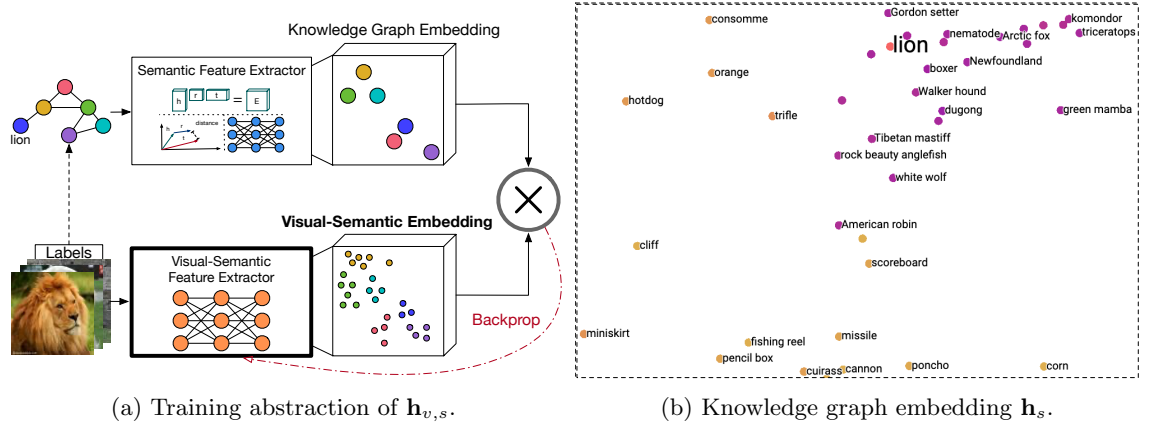


Figure 5.1: Overview of the KG-NN approach: a) the main building blocks for learning a visual-semantic embedding space $\mathbf{h}_{v,s}$ using a *Knowledge Graph as a Trainer*; b) the 2D projection of the semantic-embedding \mathbf{h}_s represented in a knowledge graph.

The chapter starts with Section 5.1, which presents a detailed description of KG-NN, the approach that uses a KG to train a DNN for CV. Section 5.2 provides an evaluation on two datasets in a domain generalization and domain adaptation task. We summarize related work in Section 5.3 and conclude the chapter and provide an outlook on future directions in Section 5.4.

5.1 Knowledge Graph as a Trainer

In this section, we define the basic terminology of the KG-NN approach as well as the underlying pipeline for the realization of a transfer learning task. The main objective of KG-NN is incorporating prior knowledge into the DL pipeline using a *KG as a Trainer*.

The main concept of KG-NN is illustrated in Figure 5.1. As depicted in Figure 5.1a, the class labels of a given dataset are infused to the DNN in form of a high-dimensional vector of the KGE \mathbf{h}_s , instead of using the standard one-hot encoded vectors. This \mathbf{h}_s shown in Figure 5.1b is generated from a KG using a KGE method $KGE(\cdot)$. It incorporates domain-invariant relations to other classes inside or outside the dataset and therefore enriches the DNN with prior knowledge in an indirect manner. To guide the adaption of the DNN to the \mathbf{h}_s space, we use the *contrastive knowledge graph embedding loss*. It compares the respective outputs from the visual feature extractor with the class label vectors of the \mathbf{h}_s forming a visual-semantic embedding space $\mathbf{h}_{v,s}$. As a result, the learned DNN projects respective images close to their representations given by the \mathbf{h}_s .

5.1.1 Knowledge Infusion

We infuse the knowledge of the KGE into the DNN via the contrastive knowledge graph embedding loss. We derive the loss from the supervised contrastive loss [250, 87] which extends the multi-class

N-pair loss [251] or InfoNCE loss [252] with class label information. Instead of contrasting images in the batch against an anchor image, we adapt the loss to contrast images of the batch against the class label representation of the \mathbf{h}_s . A batch consists of $2N$ training samples and two augmented versions for each of the N training images. Within a batch, an anchor $i \in \{1 \dots 2N\}$ is selected that corresponds to a specific class label \mathbf{y}_i and therefore assigns a specific embedding vector of the \mathbf{h}_s , $h_{s,i}$. Positive samples are all samples that correspond to the same class label as the anchor i . The numerator in the loss function computes a similarity score between the anchor vector of the \mathbf{h}_s , $h_{s,i}$, and the visual projection vector of a positive sample in the batch, $\mathbf{h}_{v,j}$. The denominator computes the similarity score between the anchor vector of the \mathbf{h}_s and the visual projection vector of all other samples $\mathbf{h}_{v,k}$ in the batch. We choose the cosine similarity as the distance measure in the high-dimensional space. For each anchor i , there can be many positive samples, which contribute to the final loss, where $N_{\mathbf{y}_i}$ is their total number. The KG-based contrastive loss function is then given by:

$$\mathcal{L}_{KG} = \sum_{i=1}^{2N} \mathcal{L}_{KG,i} \quad (5.1)$$

with

$$\mathcal{L}_{KG,i} = \frac{-1}{2N_{\mathbf{y}_i} - 1} \sum_{j=1}^{2N} \mathbb{1}_{i \neq j} \cdot \mathbb{1}_{\mathbf{y}_i = \mathbf{y}_j} \cdot \log \frac{\exp(\mathbf{h}_{s,i} \cdot \mathbf{h}_{v,j} / \tau)}{\sum_{k=1}^{2N} \mathbb{1}_{i \neq k} \exp(\mathbf{h}_{s,i} \cdot \mathbf{h}_{v,k} / \tau)} \quad (5.2)$$

where $\mathbf{h}_v = P(E(\mathbf{x}))$, $\mathbb{1}_{k \neq i} \in \{0, 1\}$ is an indicator function that returns 1 iff $k \neq i$ evaluates as true, and $\tau > 0$ is a predefined scalar temperature parameter. During optimization of the loss function \mathcal{L}_{KG} , the DNN learns its weights by mapping its projection \mathbf{h}_v to the \mathbf{h}_s space.

5.1.2 Adaptation to a Labeled Target Domain

Training robust DNNs is crucial in real-world scenarios as deployment domains typically differ from the training ones. The *KG as a Trainer* can influence how a DNN should behave in different environments by providing a stable embedding space. However, if the domain gap is quite large, it is beneficial to fine-tune the DNN on labeled data of the target domain.

We design a training pipeline to support a transfer learning scenario where a small amount of labeled target data exists. An overview of this pipeline comprised of five consecutive phases is shown in Figure 5.2.

Knowledge Graph Construction: KGs can represent prior knowledge encoded with rich semantics in a graph structure. Based on the selected scenario, underlying knowledge of one or multiple domains is conceptualized and formalized into a KG. Since KGs are manually curated by human experts, it is possible to define an underlying schema comprising multiple classes from different domains. This joint representation of domains enables inferring relations between classes, which can then be transferred into high-dimensional vector space.

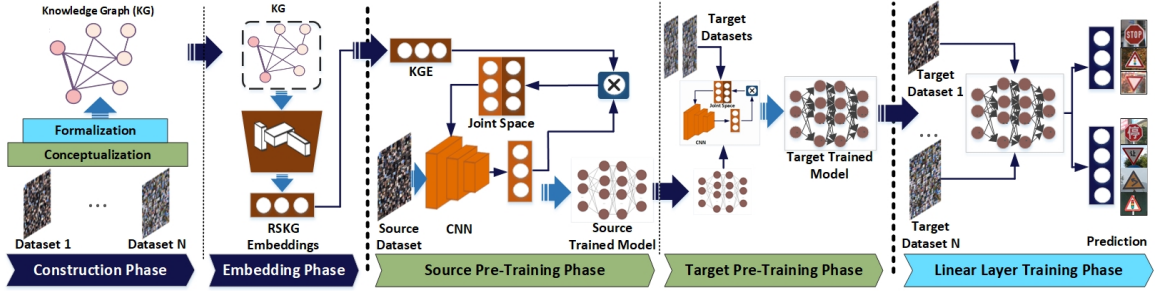


Figure 5.2: The pipeline designed for KG-NN consists of five phases in which a knowledge graph acts as a trainer that supports the adaptation and generalization of DL: *Knowledge Graph Construction*; *Knowledge Graph Embedding*; *Source Domain Pre-Training*; *Target Domain Pre-Training*; and *Linear Layer Training*.

Knowledge Graph Embedding: The KG is transformed into a KGE \mathbf{h}_s via a KGE method $KGE(\cdot)$. There are various approaches to generate these dense vectors that encode all entities and relations within the KG [253, 254, 255]. Note that KG-NN can operate on any \mathbf{h}_s generated by any $KGE(\cdot)$, as an \mathbf{h}_s only reflects similarities between entities by distances and positions in the vector space. Thus, if entities share many properties in the KG, they are closely located in space.

Source Domain Pre-Training: We train KG-NN from scratch using the *KG as a Trainer* and do not initialize the DNN with pre-trained weights from ImageNet [173]. As the $\mathbf{h}_{v,s}$ space of KG-NN depends on the KG instead of the source dataset, KG-NN can be applied to other domains following the same semantic relations given by the KG. This property is shown on the domain generalization task.

Target Domain Pre-Training: Small amounts of labeled target data can usually be gathered with manageable effort. However, just fine-tuning a DNN with additional target domain data using the cross-entropy loss leads to catastrophic forgetting and thus poor accuracy. We assume that this happens because the DNN tries to find a new \mathbf{h}_v that fits the target domain, but differs from the embedding obtained from the source domain. In contrast, DNNs optimized on the source domain using a *KG as a Trainer* can simply be enriched with additional target data using the same training method. Therefore, KG-NN pre-trained on the source domain is retrained on the target dataset using the same \mathbf{h}_s .

Linear Layer Training: For adaption to a downstream task like classification, we add a randomly-initialized linear fully-connected layer to the trained encoder network. The size of the output vector depends on the number of classes. This linear layer is trained with the default cross-entropy loss, while the parameters of the encoder network remain unchanged.

5.2 Experiments

We conduct experiments on two different scenarios with multiple datasets to demonstrate the benefit of training a DNN using a *KG as a Trainer*, which leads to more accurate and more robust models in terms of the distribution shift. We compare KG-NN with two baselines: 1) CE, which trains the DNN using the supervised cross-entropy loss; and 2) SupCon, which trains the DNN with the (self-)supervised contrastive loss [250]. We chose CE, as it is typically used for training DNNs, as well as SupCon, as this approach utilizes a similar contrastive loss function, however without the incorporation of prior knowledge and supervision. CE and SupCon learn an embedding layer based on the data distribution of the source dataset, whereas KG-NN relies on the embedding given by the KG. To qualitatively evaluate the influence of the KGE we further compare against GloVe, a variation of KG-NN that uses a GloVe [227] language embedding instead of \mathbf{h}_s . All approaches use the same ResNet-50 [256] backend as the encoder network and only differ in their method of how this encoder network is trained.

Two different scenarios are defined to analyze our approach to concrete transfer learning tasks. *Scenario 1* - we investigate the sensitivity to distribution shifts using a domain generalization task. Therefore, we train: a) KG-NN, CE, SupCon, and GloVe from scratch on Mini-ImageNet and evaluate on its derivatives, ImageNetV2 [174], ImageNet-R [257], ImageNet-Sketch [258] and ImageNet-A [259]; b) KG-NN, CE, and SupCon from scratch on GTSRB, and evaluate on CTSD. *Scenario 2* - we focus on supervised domain adaptation, a more practical scenario where KG-NN, CE, and SupCon are trained on GTSRB and fully retrained on CTSD with a small amount of target data. Note that we exclude GloVe when using GTSRB/CTSD since the language embedding does not contain a specific representation for each road sign class and therefore can not be applied straightforwardly.

5.2.1 Scenario 1 - Domain Generalization

Domain generalization describes the task of learning generalized models on a source domain so that they can be used on unseen target domains. Therefore, KG-NN is used without the target domain pre-training phase.

Experiment 1 - Wordnet-Subset with Mini-ImageNet

Dataset Settings: As the source domain, we use Mini-ImageNet, a derivative of the ImageNet dataset, consisting of 60 thousand color images of size 84×84 with 100 classes, each having 600 examples. Compared to ImageNet, this dataset fits in memory on modern machines, making it very convenient for rapid prototyping and experimentation. For the evaluation, we use the target domains: ImageNetV2, which contains 10 new test images per class and closely follows the original labeling protocol; ImageNet-R, which has art, cartoons, deviantart, etc. renditions of 200 ImageNet classes resulting in 30 thousand images; ImageNet-Sketch comprising 50 thousand images, 50 images for

each of the one thousand ImageNet classes; and ImageNet-A, which contains real-world, unmodified, and naturally occurring examples that cause ML model’s performance to significantly degrade.

Knowledge Graph and KG Embedding Space: WordNet is a lexical database containing nouns, verbs, adjectives, and adverbs of the English language structured into respective synsets [148, 260]. Each synset is an underlying concept consisting of a collection of synonyms as well as its relations to other synsets. The *mini WordNet knowledge graph* (MWKG) is created by extracting the respective synsets of each label from the Mini-ImageNet dataset from [261] into RDF representation. These synsets are grouped based on the lexical domain they pertain to, e.g. *animal*, *artifact*, or *food*. They are represented as classes and further described with relations such as: *hypernym*, *meronym*, *synset-member*. Additionally, a shallow taxonomy is established by extracting the parents of each synset including their relationships and attributes. In total, MWKG contains 198 classes with eight annotation properties. We transfer MWKG into a 300-dimensional \mathbf{h}_s using the MRGCN [262], which exploits the literal information in addition to classes and their relationships. To realize that, we use the MRGCN’s node classification feature to build the \mathbf{h}_s that explicitly clusters the six lexical domains: animal, artifact, communication, food, object, and plant.

Training Details: All models use a ResNet-50 backend and are pre-trained with a batch size of 1,024 on the Mini-ImageNet dataset. We resize the images to 32x32 for fast prototyping. KG-NN and SupCon are trained for one thousand epochs using their respective contrastive loss function, SGD with a learning rate of 0.5, cosine annealing, and a temperature of $\tau = 0.5$. CE is trained for 500 epochs with the cross-entropy loss and SGD with a learning rate of 0.8. For the *linear-layer phase*, we train an *one-layer* MLP on top of the frozen encoder networks of KG-NN, SupCon, and CE, with an adam optimizer and a learning rate of 0.0004.

Evaluation: We evaluate the models on ImageNetV2, ImageNet-R, ImageNet-Sketch, and ImageNet-A. KG-NN outperforms CE, SupCon, and GloVe on the trained source as well as on unknown target domains as shown in Figure 5.3. This means that KG-NN makes use of the additional semantic information. It can be seen that CE fails particularly when the domain gap increases. We assume that this happens due to its high specialization on the source domain. SupCon cannot reach the performance of CE on the source dataset, however, it outperforms CE on more general target tasks. We see that pre-training on a more generic self-supervised task helps the DNN to extract more general features. GloVe, the version of KG-NN that relies on a language embedding instead of a KG, is also outperformed by KG-NN. We see that the performance of KG-NN depends on the quality of the embedding space, which we can control manually using different KGs or KGE methods.






		CE	SupCon	GloVe	KG-NN
Mini-ImageNet		64.5 +1.7%	56.2 +10.0%	64.7 +1.5%	66.2
ImageNetV2		49.3 +3.2%	43.8 +8.7%	49.5 +3.0%	52.5
ImageNet-R		20.6 +9.4%	26.0 +4.0%	27.4 +2.6%	30.0
ImageNet-Sketch		12.3 +15.4%	22.3 +5.4%	24.3 +3.4%	27.7
ImageNet-A		4.3 +1.0%	4.2 +1.1%	5.0 +0.3%	5.3

Figure 5.3: Accuracy of the domain generalization scenario using Mini-ImageNet as the source and multiple derivatives as target domains. We compare KG-NN with the standard CE, SupCon, a version of our loss without prior knowledge of a KG, and GloVe, a version of KG-NN using a language embedding instead of a \mathbf{h}_s .

Experiment 2 - RoadSign KG with GTSRB and CTSD

Dataset Settings: The GTSRB, which contains 51,970 images of 43 road signs, is used as the source domain, and the CTSD, which contains 6,164 images of 58 road signs, as the target domain. We resize all images to a uniform size of 32×32 pixels. Note that we do not cut out the road signs, but take the whole image for classification. Both datasets contain a domain shift as they were recorded with different cameras in different countries and hence have different appearances.

Knowledge Graph and KG Embedding Space: First, we construct a small KG for traffic sign recognition, the *road sign knowledge graph* (RSKG) that contains all classes of both datasets incorporated in an underlying domain ontology.

An example subgraph of RSKG is illustrated in Figure 5.4. To encode the formal semantics of road signs from different countries and standards, we first develop the *RoadSign* ontology. It contains classes (e.g. RoadSign, Shape, Icon, Color), relationships (e.g. hasShape, hasIcon, hasColor), and attributes (e.g. label, textWithinSign, etc). The actual road signs that exist within given datasets are represented as concrete *individuals*. Note that this information is extracted from externally available road-sign standards, without accessing the datasets. Currently, RSKG contains 18 classes, 11 object properties, two datatype properties, and 101 individuals.

It is important to mention that the KG can be further populated with concrete road signs instances from other countries. This would enrich RSKG and could help to find inter-relations between the domains. We transfer RSKG into a 300-dimensional \mathbf{h}_s by using MRGCN [262] as we also want to exploit its literal information. Therefore, we use MRGCN in the node classification task

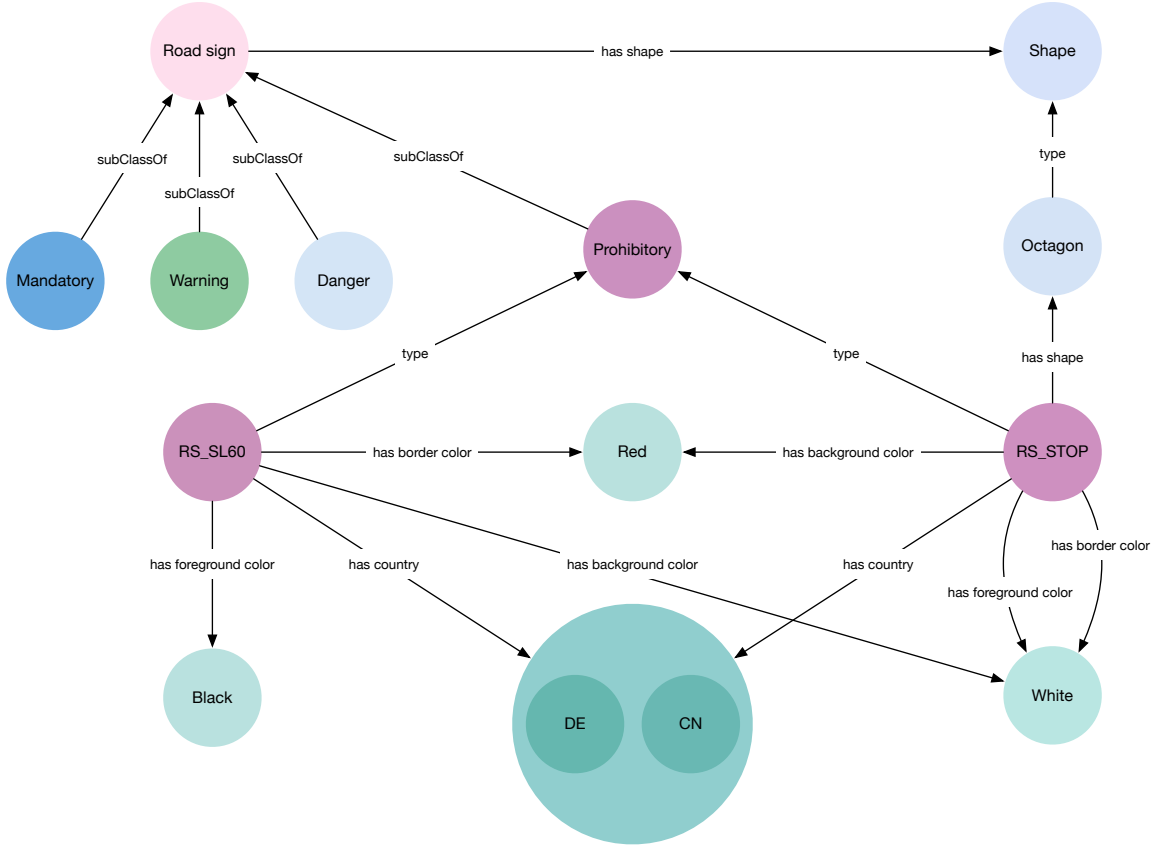


Figure 5.4: The *road sign knowledge graph* (RSKG) models the class-based relationships of the individual road signs of the GTSRB and CTSD datasets. It includes visual properties, such as shape, foreground color, etc., but also features based on road sign conventions, e.g. type, and country information.

to build a KGE. The KGE is therefore trained to clusters the \mathbf{h}_s into the five subclasses of danger, informative, mandatory, prohibitory, and warning, given their underlying other features in the KG.

Training Details: All models use a ResNet-50 backend and are pre-trained with a batch size of 1,024 on the GTSRB dataset. We resize the images to 32x32 for fast prototyping. KG-NN and SupCon are trained for one thousand epochs using their respective contrastive loss function, SGD with a learning rate of 0.5, cosine annealing, and a temperature of $\tau = 0.5$. CE is trained for 500 epochs with the cross-entropy loss and SGD with a learning rate of 0.8. For the *linear-layer phase*, we train an *one-layer* MLP on top of the frozen encoder networks of KG-NN, SupCon, and CE, with an adam optimizer and a learning rate of 0.0004.

		CE	SupCon	KG-NN
GTSRB		96.1 +0.8%	41.9 +55.0%	96.9
CTSD		63.0 +7.1%	34.4 +35.7%	70.1

Figure 5.5: Accuracy of the domain generalization scenario using GTSRB as the source and CTSD as the target domain. We compare KG-NN with the standard CE and SupCon, a version of our loss without prior knowledge of a KG.

Evaluation: Figure 5.5 shows that KG-NN outperforms CE by 0.8% on the source and by 7.1% on the target dataset. It can be seen that KG-NN exceeds the accuracy of SupCon by 55.0% on GTSRB and by 35.7% on CTSD. SupCon with its self-supervised loss needs large datasets to form a good embedding space, however, both datasets are quite small and from the special domain of road sign recognition. We do not compare against a GloVe embedding, as there are no instances for specific road signs and no clear procedure on how to generate these instances from a text corpus. Overall, KG-NN performs better and is more robust to unforeseen distribution shifts using the same amount of training data.

5.2.2 Scenario 2 - Supervised Domain Adaptation

Supervised domain adaptation describes the task of transfer learning that adapts models learned on a source domain to a specific labeled target domain. We claim that a DNN learned using an image-data-independent \mathbf{h}_s can adapt to new domains and new classes better as both domains use the same embedding space. For this experiment, we use the settings described in Experiment 2.

First, KG-NN, CE, and SupCon are pre-trained on the source dataset. Second, we use the encoder networks of each DNN and presume the pre-training on the target dataset. The DNNs are retrained with different amounts of labeled target data. The one-shot (58) experiment uses 58 images, one image for each class of the CTSD target dataset. The five-shot (290) experiment uses 290 images, five images for each class of the CTSD. The 10% (416) experiment uses 416 images, 10% of images of the CTSD. The 50% (2083) experiment uses 2,083 images, 50% of images of the CTSD. The 100% (4165) experiment uses 4,165 images, 100% of images of the CTSD target dataset.

Similar to the previous experiments, we use the *linear layer phase* to adopt the pre-trained encoder network to the target task. As shown in Figure 5.6, all experiments are evaluated on the full CTSD target dataset and on the 25 common classes of the GTSRB source dataset.

Evaluating the approaches on the initial source domain, we find that all DNNs suffer from *catastrophic forgetting*, as depicted in Figure 5.6b. If 100% of target data is used for training, the accuracy of CE drops from 96.1% to 49.5%, the accuracy of SupCon drops from 41.9% to 37.2%, and

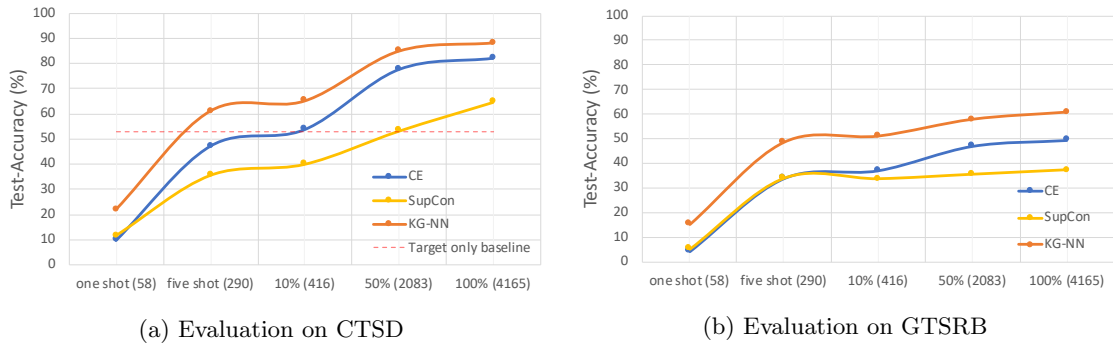


Figure 5.6: Evaluation of the domain adaptation scenario for GTSRB to CTSD. We compare GTSRB-trained KG-NN, SupCon, and CE for domain adaptation when re-trained with different amount of target data from CTSD: a) evaluates the DNNs on the target domain (CTSD); b) evaluates the DNNs again on the initial source domain (GTSRB) to reflect the *catastrophic forgetting* phenomenon.

the accuracy of KG-NN drops from 96.9% to 60.7% on the source domain. This means that KG-NN is still the best-performing model on the source domain, even after retraining on a target domain with an increased difference to CE from 0.8% to 11.2%. We think that the fixed embedding space between the source and the target domain helps to overcome the issue of *catastrophic forgetting*.

If we compare the approaches on the target domain as illustrated in Figure 5.6a, we see that KG-NN achieves an accuracy of 88.1%, which is an improvement by 5.9% over standard CE and by 23.4% over SupCon. Since we operate on transfer learning, an additional target-only baseline is introduced. Thus, CE is initialized with weights pre-trained on ImageNet, instead of using the source domain to pre-train the parameters of the DNN. We see that the target-only baseline suffers from fewer target data in D_T yielding only 53.1% accuracy as the ImageNet initialization does not suit well for the task of road sign recognition. All approaches seem to be able to transfer some knowledge from the source domain D_S to the target domain D_T outperforming the target-only baseline. However, KG-NN significantly outperforms the baseline by 35.0%, whereas CE improves by 29.1% and SupCon by 11.6%.

Interestingly, with less than five target images per class, which is fewer than 7% of target data, KG-NN surpasses the performance of the target-only baseline. We observe KG-NN always outperforms CE by approximately 10% of accuracy. When compared to SupCon, we see the accuracy difference even increases if more labeled target data is available. In the one-shot scenario, KG-NN outperforms CE by 12.2% of accuracy, in the five-shot-scenario by 13.8%, in the 10%-scenario by 11.2%, in the 50%-scenario by 10.7%, and on the full target dataset by 5.9%. In the one-shot scenario, KG-NN outperforms SupCon by 10.3% of accuracy, in the five-shot-scenario by 25.4%, in the 10%-scenario by 25%, in the 50%-scenario by 31.6%, and on the full target dataset by 23.4%.

5.3 Work Related to Learning Visual Models using a KG

Embedding spaces trained with the cross-entropy loss tend to be specialized embedding spaces for a particular domain. To reduce the high dependency on the training domain, pre-training methods that generate rich embedding spaces seem to be a promising research direction for CV and NLP. Most neuro-symbolic approaches only learn a transformation function, e.g., MLP, on top of a pre-trained \mathbf{h}_v . We refer to these models as visual-semantic transformation models. Since the weights of the visual feature extractor are a really important part of robust object recognition, recent approaches have shown that learning a visual-semantic feature extractor from scratch improves generalization capabilities and makes the DNN applicable to further downstream and transfer learning tasks [86]. We refer to these models as visual-semantic features extractors.

Neural Networks improved by KG: Most of the works that combine KGs with NNs use WordNet [263], small-scale label [133, 204] or scene [195] graphs as KG. However, the capacity of WordNet as a lexical database is limited. Large-scale KGs such as DBPedia [149] or ConceptNet [152] encode additional semantic information by using higher order relations between concepts. Although their applications are still sparse in the visual domain, there are a few works that have shown promising results. DBPedia is already used in the field of explainable AI [264, 265], object detection [136], and visual question answering [146]; and ConceptNet is used for video classification [145] and zero-shot action recognition [137]. However, all approaches use the KG only as a post-validation step on a pre-trained visual feature extractor, while KG-NN learns the visual feature extractor by itself based on the KG.

Visual-Semantic Transformation Models: Visual-semantic transformation models are learned via a transformation function, e.g. MLP, from a pre-trained \mathbf{h}_v into \mathbf{h}_s . One of the first approaches that use \mathbf{h}_s with NNs is the work from Mitchell et al. [209]. They use word embeddings derived from text corpus statistics to generate neural activity patterns, i.e. images. Instead of generating images from text, Palatucci et al. [205] learn a linear regression model to map neural activity patterns into word embedding space. In their work, they improve zero-shot learning by extrapolating the knowledge gathered from in the \mathbf{h}_s related classes to novel classes. Socher et al. [96] present a model for zero-shot learning that learns a transformation function between an \mathbf{h}_v space, obtained by an unsupervised feature extraction method, and an \mathbf{h}_s , based on an NN-based language model. The authors trained a 2-layer NN with the MSE loss to transform the \mathbf{h}_v into the word embedding of eight classes. Frome et al. [99] introduce the deep visual-semantic embedding model DeVISE that extends the approach from eight known and two unknown classes to one thousand known classes for the image model and up to 20 thousand unknown classes. Therefore, they pre-train their visual feature extractor using ImageNet and their \mathbf{h}_s based on the Word2Vec [230] language model, exposed to the text of a single online encyclopedia. In contrast to Socher et al. [96], DeVISE learns

a linear transformation function between the \mathbf{h}_v space and the \mathbf{h}_s space using a combination of dot-product similarity and hinge rank loss since MSE distance fails in high-dimensional space. Norouzi et al. [206] propose *convex combination of semantic embeddings* (ConSE), a simple framework for constructing a zero-shot learning classifier. ConSE uses a semantic word embedding model to reason about the predicted output scores of the NN-based image classifier. To predict unknown classes, it performs a convex combination of the classes in the \mathbf{h}_s space, weighted by their predicted output scores of the NN. Similarly, Zhang et al. [207] introduce the *semantic similarity embedding* (SSE), which models target data instances as a mixture of seen class proportions. SSE builds a semantic space where each novel class could be represented as a probabilistic mixture of the projected source attribute vectors of the seen classes. Akata et al. [160] refer to their \mathbf{h}_s space transformations as label embedding methods. They compared transformation functions from the \mathbf{h}_v space to the attribute label embedding space, the hierarchy label embedding space, and the Word2Vec label embedding space, in which embedded classes can share features among themselves.

Visual-Semantic Features Extractors: The approaches mentioned so far only learn a transformation from \mathbf{h}_v to \mathbf{h}_s . However, the parameters of the feature extractor are not affected by the prior information. Thus, if the feature extractor cannot detect visual features due to the domain shift problem, the performance of the final prediction suffers. Instead of maximizing the likelihood on the output, some approaches maximize the energy, i.e. the difference between the prediction and the expected result, directly on the embedding space to learn the NN. Hadsell et al. [266] introduce the contrastive loss for a *siamese architecture* to learn a robust embedding space from unlabeled data. They show that their self-supervised energy-based method can learn a lighting and rotation-invariant embedding space. Recently, many approaches claim that training an embedding space in a self-supervised manner using the contrastive loss tends to find a more general and domain-invariant representation [87, 267]. Furthermore, Tian et al. [268] show that learning an embedding space using the contrastive loss, followed by training a supervised linear classifier on top of this representation, outperforms state-of-the-art few-shot learning methods.

Joulin et al. [212] demonstrate that feature extractors trained to predict words in image captions can learn useful visual-semantic embedding spaces $\mathbf{h}_{v(s)}$. Further, Radford et al. [86] proposed a general pre-training of an NN with natural language supervision using a dataset of 400 million image-text pairs collected from the Internet and the contrastive objective of Zhang et al. [225].

To the best of our knowledge, there is no work that learns a visual feature extractor using a KG or its embedding space \mathbf{h}_s . We choose to use prior knowledge encoded in a KG instead of using the unstructured knowledge of a language embedding as they are highly dependent on their text corpus, inconsistent, and do not incorporate expert knowledge.

5.4 Summary

In this Chapter, we proposed KG-NN, a KG-based approach that enables DNNs to learn more robust and controlled embedding spaces for visual transfer learning. The core idea of our approach is to use domain-invariant knowledge represented in a KG, transform it into a vector space using a KGE method, and train a DNN so that its embedding space is adapted to the domain-invariant embeddings given by the KG. The KG relies on prior knowledge of the domain and related domains represented in the form of symbolic graph structures. A domain-invariant \mathbf{h}_s is formed by transforming the symbolic knowledge of the KG with KGE methods into a high-dimensional vector space. Using our KG-based contrastive loss function, we force the DNN to adapt its \mathbf{h}_v space to the domain-invariant \mathbf{h}_s space given by the KG, thus forming $\mathbf{h}_{v,s}$. Our experimental results show that DNNs benefit from exploiting prior knowledge, in particular on visual transfer learning tasks. As a result, KG-NN increases the generalization performance of DNNs and therefore the accuracy on known and unknown domains. For domain adaptation tasks KG-NN keeps up with DNNs trained with the cross-entropy loss despite requiring significantly less training data.

Chapter 6

Context-driven Visual Object Recognition based on Knowledge Graphs

How humans perceive the real world is strongly dependent on the context [269, 270]. Especially, in situations with poor quality of visual input, for instance, caused by large distances, or short capturing times, context appears to play a major role in improving the reliability of recognition [271]. Perception is not only influenced by co-occurring objects or visual features in the same image, but also by experience and memory [272]. There is evidence that humans perceive similar images differently considering the given context [273]. A famous example are ambiguous figures as shown in Figure 6.1.

Depending on the context, i.e. if it is Easter or Christmas [276], Figure 6.1a can be either a duck or a rabbit. Likewise, influenced by own-age social biases [277], Figure 6.1b can be either a young lady or an old woman.



(a) Duck or rabbit? [274]



(b) Young lady or old woman? [275].

Figure 6.1: Humans perceive similar images differently considering the given context. Ambiguous figures show that the perception and mental representation for visually similar input can change depending on the context.

Humans categorize images based on various types of context. Known categories are based on visual features or semantic concepts [278], but may also be based on other information such as attributes describing their function. Accordingly, neuroscience has shown that the human brain encodes visual input into individual contextual object representations [5, 6, 7], namely visual, taxonomical, and functional [279]. Concretely, in a visual context, images of a drum and a barrel have a high similarity, as they share similar visual features. In a taxonomical context, a drum would be similar to a violin, as they both are musical instruments. And in a functional context, the drum would be similar to a hammer, since the same action of hitting can be performed with both objects [280].

Whereas there is much evidence that intelligent machines should also represent information in contextualized embeddings, deep neural networks form their object representations based only on the feature distribution of the image dataset [281, 282]. Therefore, they fail if the objects are placed in an incongruent context that was not present in previously seen images [283].

This chapter investigates the influence of the type of prior knowledge of the KG that is induced into the DNN. Does taxonomical prior knowledge, equally benefit visual object recognition, as functional or visual prior knowledge? What is the relationship between prior knowledge, the image data, and the task? To investigate these questions, we propose a theoretical framework that extracts contextual views of a KG, which then are used to train KG-NNs individually.

In this chapter, we address the following research question:

RQ4: How does the type of prior knowledge in a KG impact KG-DL performance, especially in visual transfer learning?

To provide a comprehensive answer for **RQ4**, we divided the research question into:

- **RQ4.1:** Can context provided in the form of a KG influence learning image representations of a DNN, the final accuracy, and the image predictions?
- **RQ4.2:** Can context help to avoid critical errors in domain-changing scenarios where DNNs fail?

Our main contribution based on **RQ4** is as follows:

- *Method to learn contextual DNNs using contextual views of a generic KG.*

The main parts of this chapter are already published in the following work:

- **Sebastian Monka**, Lavdim Halilaj, and Achim Rettinger. 2022. Context-Driven Visual Object Recognition Based on Knowledge Graphs. In *The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings (Lecture Notes in Computer Science)*, Springer, 142–160.

To enable standard DNNs to build contextual object representations, we provide the context using a KG and its corresponding KGE \mathbf{h}_s . Similar to the process in the human brain, we conduct experiments with three different types of contexts, namely visual context, taxonomical context, and functional context 6.3. We provide two versions of knowledge infusion into a DNN and compare the induction of different contextual models in depth by quantitatively investigating their learned contextual embedding spaces using class-related cosine similarities. In addition, we evaluate our approach quantitatively by comparing their final accuracy on object recognition tasks on source and target domains and provide insights and challenges.

The structure of this chapter is organized as follows: Section 6.2.1 introduces the three different types of context and an option to model these views in a contextual KG. Section 6.2 shows two ways of infusing context into a visual DNN. Section 6.3 contains experiments on seven image datasets within two transfer learning scenarios. Section 6.4 provides an overview of work related to the content of this chapter. Section 6.5 answers the research questions and summarizes our main insights.

6.1 Contextual Image Representations

Contextual Image Representations in the Brain: Cognitive and neuroscience research has recently begun to investigate the relationship between viewed objects and the corresponding fMRI scan activities of the human brain. It is assumed that the primate visual system is organized into two separate processing pathways in the visual cortex, namely, the *dorsal pathway* and the *ventral pathway*. While the dorsal pathway is responsible for the spatial recognition of objects as well as actions and manipulations such as grasping, the ventral pathway is responsible for recognizing the type of object based on its form or motion [284]. Bonner et al. [285] recently showed that the sensory coding of objects in the ventral cortex of the human brain is related to statistical embeddings of object or word co-occurrences. Moreover, these object representations potentially reflect a number of different properties, which together are considered to form an object concept [279]. It can be learned based on the context in which the object is seen. For example, an object concept may include the visual features, its taxonomy, or the function of the object [7, 6].

Image Representations in the DNN: Recent work has shown that while the performance of humans, monkeys, and DNNs is quite similar for object-level confusions, the image-level performance does not match between different domains [7]. In contrast to visual object representations in the brain, which also include high-level contextual knowledge of concepts and their functions, image representations of DNNs only depend on the statistical co-occurrence of visual features and a specific task. We consider the context extracted from the dataset as dataset bias. Even in balanced datasets, i.e., datasets containing the same number of images for each class, there still exists an imbalance due to the overlap of features between different classes. For instance, it must be taken into account

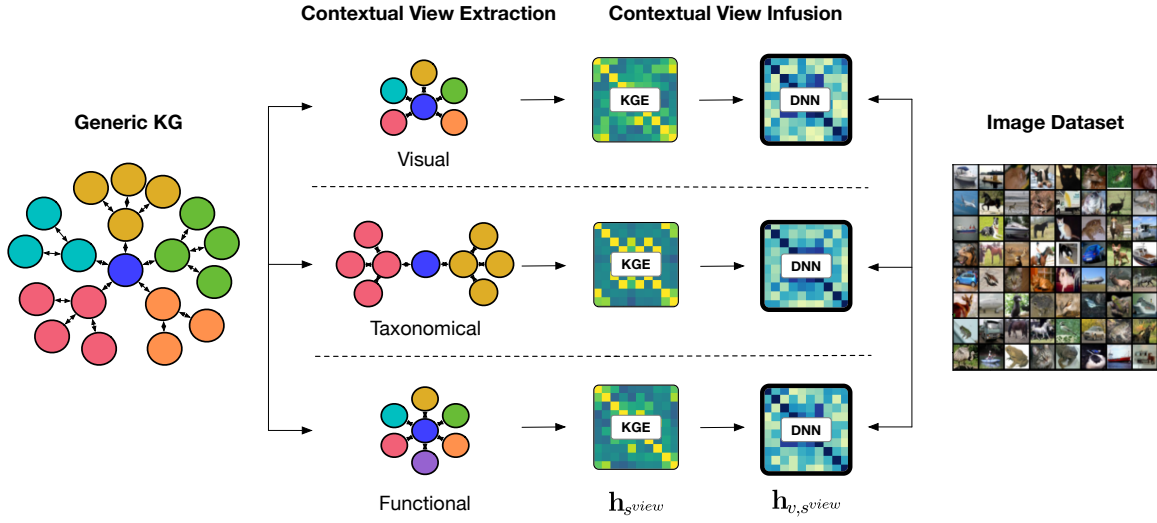


Figure 6.2: A Framework to learn contextual *deep neural networks* (DNN) using contextual views provided by a generic knowledge graph. Our approach to learning contextual image representations consists of two main parts: 1) the *contextual view extraction*; and 2) the *contextual view infusion*.

that a cat and a dog have similar visual features and that in composite datasets certain classes can have different metadata for the images, such as illumination, perspective, or sensor resolution. This dataset bias leads to predefined neighborhoods in the visual embedding space, as well as predefined similarities between distinct classes. In a DNN, a feature extractor $f(\cdot)$ maps images \mathbf{x} to a visual embedding $\mathbf{h}_v = f(\mathbf{x}) \in \mathbb{R}^{d_E}$, where the activations of the final pooling layer and thus the representation layer have a dimensionality d_E , where d_E depends on the encoder network itself.

Contextual Representations in the KG: A KG is a graph of data aiming to accumulate and convey real-world knowledge, where entities are represented by nodes and relationships between entities are represented by edges [41]. A generic KG GKG is a graph of data that relates different classes of a dataset based on defined contextual properties. These contextual properties can be both learned and manually curated. They bring in prior knowledge about classes, even those that may not necessarily be present in the image dataset, and thus place them in contextual relationships with each other.

6.2 Learning Contextual Image Representations

The framework, as shown in Figure 6.2 consists of two main parts: 1) the *contextual view extraction*, where task-relevant knowledge is extracted from a GKG ; and 2) the *contextual view infusion*, where the contextual view is infused into the DNN.

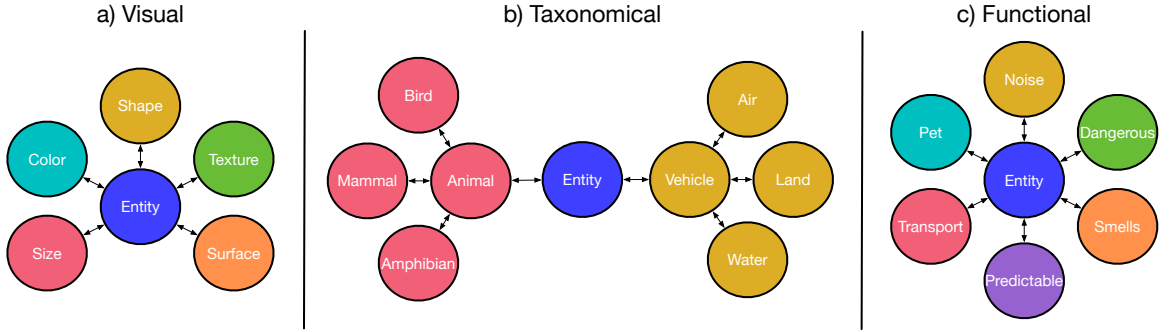


Figure 6.3: There are different types of context. Aligned to insights into how humans perceive the world, we present three contextual views of a generic knowledge graph, namely the visual, taxonomical, and functional view.

6.2.1 Contextual View Extraction

A KG can represent prior knowledge encoded with rich semantics in a graph structure. A GKG encapsulating n contextual views:

$$GKG \supseteq \{GKG^1, GKG^2, \dots, GKG^n\}$$

is a collection of heterogeneous knowledge sources, where each contextual view defines specific relationships between encoded classes. However, for a particular task, only a specific part of a GKG can be relevant. Thus, a subgraph containing a single contextual view:

$$GKG^{view} = query(GKG; view)$$

or a combination of views is extracted from a GKG . Since object recognition models are deployed in the real world that differs from their training domain, it is necessary to encode prior knowledge that is not present in the dataset.

Based on image representations in our brain and on how humans tend to classify objects, we introduce three distinct types of contextual views as shown in Figure 6.3. The first contextual view is based on visual, the second view is based on taxonomical, and the third view is based on functional properties.

Visual Context: The visual view GKG^v describes high-level visual properties of the classes, for instance, properties describing color, shape, or texture. These properties may or may not be present in the image data set. For example, if all horses in the dataset are white, we want to encode that horses can also occur in different colors.

Taxonomical Context: The taxonomical view GKG^t describes class relationships based on hierarchical schemes. A taxonomy is built by experts and can contain categories based on concepts from biology, living place, feeding method, etc. For instance, a biological taxonomy separates animals from vehicles and divides them into further subcategories.

Functional Context: The functional view GKG^f contains properties describing the function of a class. It is known that tools are categorized in the human brain based on their function [279]. In that sense properties such as hit, rub, or drill would determine the category of a given tool. However, to broaden the scope, additional functional properties such as noise, transport, or smell can be introduced.

6.2.2 Contextual View Infusion

When transferring the knowledge from the GKG^{view} using the KGE method $KGE(\cdot)$ into the KGE

$$\mathbf{h}_{s^{view}} = KGE(GKG^{view}),$$

graph-based relationships are transferred into spatial relationships. Intuitively, a different context leads to a different representation in the vector space, where $\mathbf{h}_{s^{view}}$ reflects all relationships that are modeled in GKG^{view} .

As illustrated in Figure 6.4, we present two different ways of learning a visual context embedding $\mathbf{h}_{v,s^{view}}$ in alignment with Chapter 4. The first one is KG-NN $_{u^{view}}$, which uses the *Knowledge Graph as a Trainer* 5 and thus learns $\mathbf{h}_{s^{view}}$ without any supervision of image data. The second version is KG-NN $_{s^{view}}$, which uses the *Knowledge Graph as a Peer* and thus learns $\mathbf{h}_{v,s^{view}}$ and $\mathbf{h}_{s^{view}}$ jointly with additional supervision of image data.

Both versions use the contrastive loss to align the image embedding $\mathbf{h}_{v,s^{view}}$ of the images \mathbf{x} and the DNN with the KGE $\mathbf{h}_{s^{view}}$ of the label information. A batch consists of N augmented training samples. The KG-based contrastive loss is constructed using the individual anchor losses as given by

$$\mathcal{L}_{KG^{view}} = \sum_{i=1}^N \mathcal{L}_{KG^{view},i}.$$

Within a batch, an anchor image $i \in \{1 \dots 2N\}$ is selected that corresponds to a specific class label \mathbf{y}_i , where \mathbf{y}_i points to its KGE $\mathbf{h}_{s^{view},i}$. Positive images j are all images of the batch that correspond to the same class label as the anchor i . The numerator in the loss function computes a similarity score between $\mathbf{h}_{s^{view},i}$ and the image embeddings $\mathbf{h}_{v,j}$. The denominator computes the similarity score between $\mathbf{h}_{s^{view},i}$ and the image embeddings $\mathbf{h}_{v,k}$ of all images of the other classes in the batch. As a similarity score, we choose the cosine similarity, which however can be replaced by others. $\mathbb{1}_{k \neq i} \in \{0, 1\}$ is an indicator function that returns 1 iff $k \neq i$ evaluates as true, and $\tau > 0$ is

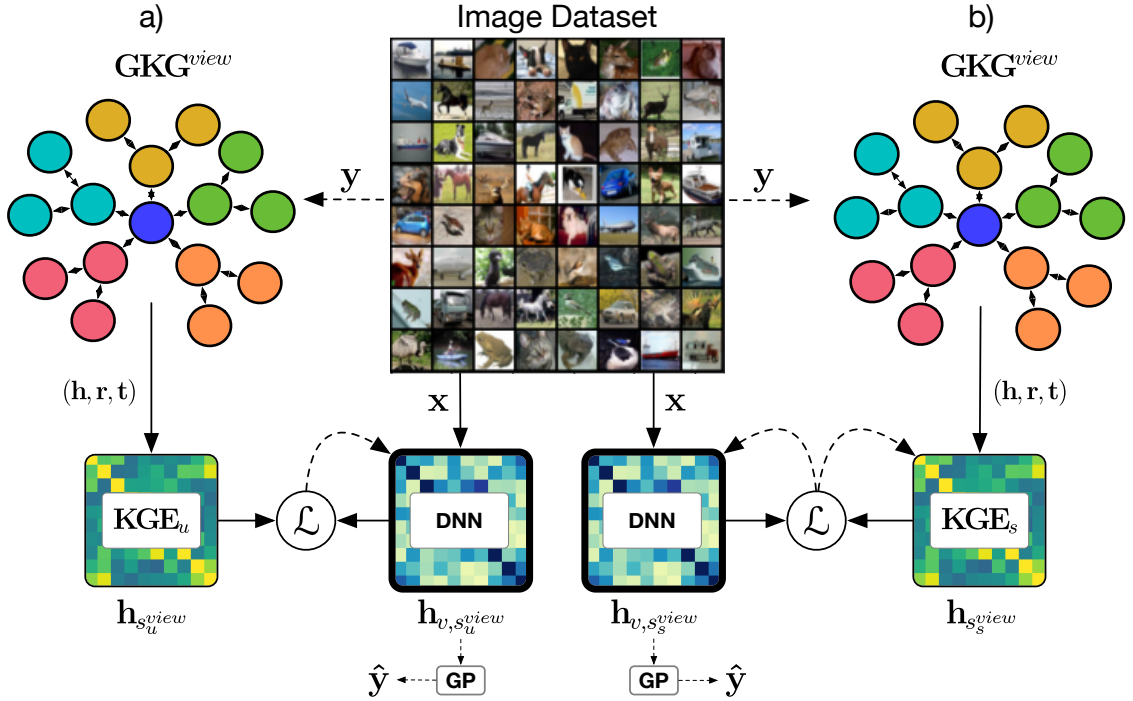


Figure 6.4: Contextual view infusion. The contextual object recognition model, i.e. *deep neural network* (DNN), can be trained in two different ways: a) using the *KG as a Trainer*, where the *knowledge graph embedding* KGE_u is learned without supervision of the image data; or b) using the *KG as a Peer*, where KGE_s is learned with supervision of the image data. Images \mathbf{x} are fed into the DNN, producing \mathbf{h}_v which is optimized with $\mathbf{h}_{s^{view}}$ using the KG-based contrastive loss and produces then $\mathbf{h}_{v,s^{view}}$. In a second step, a *gaussian process* (GP) or linear layer (LL) is trained to predict the class labels \mathbf{y} of \mathbf{x} based on the trained $\mathbf{h}_{v,s^{view}}$.

a predefined scalar temperature parameter. The KG-based contrastive anchor losses for contextual views are then given by

$$\mathcal{L}_{KG^{view},i} = \frac{-1}{2N_{\mathbf{y}_i} - 1} \sum_{j=1}^{2N} \mathbf{1}_{i \neq j} \cdot \mathbf{1}_{\mathbf{y}_i = \mathbf{y}_j} \cdot \log \frac{\exp(\mathbf{h}_{s^{view},i} \cdot \mathbf{h}_{v,j} / \tau)}{\sum_{k=1}^{2N} \mathbf{1}_{i \neq k} \exp(\mathbf{h}_{s^{view},i} \cdot \mathbf{h}_{v,k} / \tau)}.$$

Prediction: To predict the class labels of unknown images it is common to train a linear layer or to use a gaussian process on top of $\mathbf{h}_{v,s^{view}}$. For GP, we run the whole training dataset through the trained DNN and calculate the mean and covariance matrices for all the classes in $\mathbf{h}_{v,s^{view}}$. GP and LL, both calculate decision boundaries in $\mathbf{h}_{v,s^{view}}$ for all the classes of the dataset. At inference, where the goal is to predict the class label of an unknown image, GP or LL assign probabilities if an image belongs to a specific class. The maximal probability is chosen to be the final prediction.

6.3 Experiments

The goal of our empirical investigations is to provide an answer to **RQ4** and therefore to **RQ4.1** and **RQ4.2**. We conduct experiments with seven datasets in the two specific domain generalization settings, Cifar10 and Mini-ImageNet. For both experiments, we build separate GKGs that include three different contextual views, the visual (GKG^v), the taxonomical (GKG^t), and the functional (GKG^f) view, respectively. Based on the framework in Section 6.2, we use GKG^{view} to learn a contextual DNN in combination with image data. We evaluate and compare both versions of our approach, $KG\text{-}NN_{u^{view}}$ and $KG\text{-}NN_{s^{view}}$.

6.3.1 Implementation details

For both experiments, we use a similar implementation of our approach. From the GKG , we extract various GKG^{view} s using respective SPARQL queries. A ResNet-18 architecture is used as a DNN backend, with a 128-dimensional MLP as the head. We train all configurations using an ADAM optimizer, a learning rate of 0.001, no weight decay, and a cosine annealing scheduler with a learning decay rate of 0.1. The images are augmented via random cropping, random horizontal flipping, color jittering, random grayscaling, and resizing to 32x32 pixels. All models are trained for 500 epochs. For a) $KG\text{-}NN_{u^{view}}$ we transform GKG^{view} into vector space using a *graph auto encoder* (GAE) [286], which we denote as the $KG\text{-}NN_{u^{view}}$ model due to its unsupervised nature. Our GAE comprises two convolutional layers, with a hidden layer dimension of 128. We train the GAE using an ADAM optimizer with a learning rate of 0.01 for 500 epochs. For b) $KG\text{-}NN_{s^{view}}$, a *graph attention network* (GAT) [59] is trained in combination with the image data, denoted as the $KG\text{-}NN_{s^{view}}$ model due to its supervised nature. The GAT consists of two GAT-layers with 256 hidden dimensions, eight heads, and an output dimension of 128. Training is performed via the same KG-based contrastive loss from the images in addition to the GKG^{view} input. We optimize the GAT using an ADAM optimizer with a learning rate of 0.001 and no weight decay.

6.3.2 Experiments on Cifar10

Dataset Settings: The source domain Cifar10 [287] consists of six thousand 32x32 color images for each of the 10 classes, namely airplane, bird, automobile, cat, deer, dog, horse, frog, ship, and truck. The target domain St110 [288] includes 500 96x96 color images for each of the 10 classes, namely airplane, bird, automobile, cat, deer, dog, horse, monkey, ship, and truck.

Knowledge Graph Construction: We build a GKG that includes the previously discussed three types of context, namely visual, taxonomical, and functional. GKG^v contains visual properties like: *hasBackground*: air, forest, water; *hasColor*: black, blue, brown; *hasPart*: eyes, legs, wings; *hasShape*: rectangular, ellipsoid, cross; *hasSize*: large, medium, small; or *hasTexture*: dotted, striped,

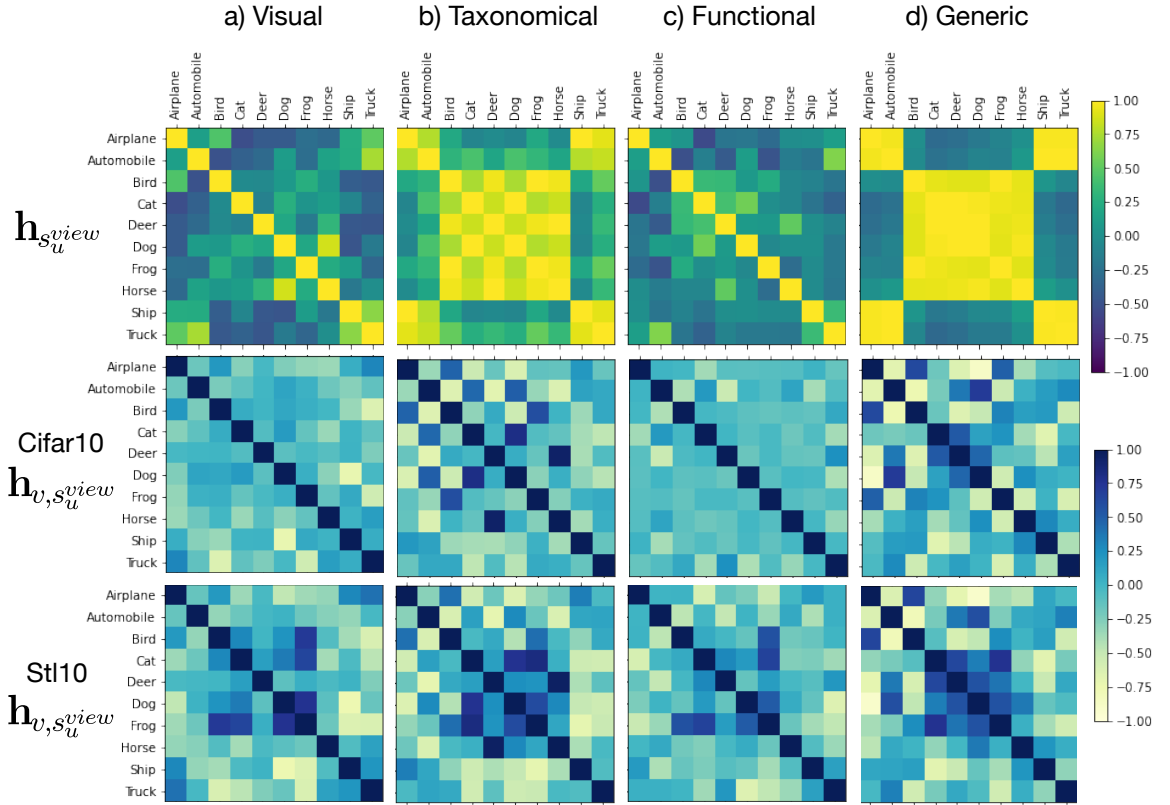


Figure 6.5: Qualitative evaluation for Cifar10. We compare $\mathbf{h}_{s_u}^{view}$ and $\mathbf{h}_{v,s_u}^{view}$ based on: a) the visual view; b) the taxonomical view; c) the functional view; and d) the full generic KG. To investigate how the semantic relationships are reflected in the embeddings, we illustrate the individual cosine similarity matrices between the classes of the Cifar10 and the Stl10 dataset.

uniform. GKG^t contains a taxonomy of the classes using the type-relation. For example, the class *Horse* is-a *Mammal* and is-an *Animal* or the class *Ship* is-a *Water-vehicle* and is-a *Vehicle*. GKG^f defines the function of the class, e.g. properties like: *hasMovement*: drive, fly, swim; *hasSound*: bark, meow, vroom; *hasSpeed*: fast, medium, slow; *hasWeight*: heavy, light, middle. Our GKG contains in total 34 classes, 16 object properties, and 65 individuals. Please note that our GKG is only an example and we are aware that there are unlimited possibilities of how and what type of knowledge can be modeled in a KG.

Evaluation: To evaluate our approach we first investigate the learned embeddings, if and how semantic relationships from GKG^{view} are reflected in $\mathbf{h}_{s_u}^{view}$. Second, we compare the individual class accuracies to see how these relationships influence the final object recognition. Figure 6.5 shows an analysis: a) the visual view; b) the taxonomical view; and c) the functional view. For every cell in $\mathbf{h}_{s_u}^{view}$ we calculate the cosine similarity between the corresponding nodes, i.e. the classes of the

(a) Results on Cifar10

Cifar10	Airplane	Auto	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	All
SupSSL	95.1	97.0	91.8	83.9	92.9	85.7	96.0	93.5	96.8	95.9	92.9
$\Delta\text{KG-NN}_{uv}$	-1.2	0.5	-2.6	-0.2	2.3	-0.8	-0.2	-1.1	-0.5	-0.9	-0.5
$\Delta\text{KG-NN}_{ut}$	-0.9	-0.6	-1.2	-30.8	-29.8	-0.2	-2.2	-1.6	-1.5	-1.3	-7.0
$\Delta\text{KG-NN}_{uf}$	1.0	0.2	-1.1	1.9	0.1	0.6	0.7	1.2	-0.1	-0.4	0.4
$\Delta\text{KG-NN}_u$	-0.7	0.0	-2.3	0.4	0.6	-0.6	-1.1	0.0	0.3	-1.8	-0.5
$\Delta\text{KG-NN}_{sv}$	-0.6	-0.3	0.2	0.3	0.1	-1.0	0.3	0.9	0.7	-0.8	-0.0
$\Delta\text{KG-NN}_{st}$	-0.9	0.0	-1.7	1.8	0.1	1.0	0.4	0.5	-0.1	0.3	0.1
$\Delta\text{KG-NN}_{sf}$	-0.4	0.5	-3.0	1.7	1.5	-0.4	0.4	0.8	0.4	-0.1	0.1
$\Delta\text{KG-NN}_s$	-1.0	0.3	-1.8	1.2	-0.3	2.0	-0.5	1.7	0.0	0.7	0.2

(b) Results on Stl10

Stl10	Airplane	Auto	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	All
SupSSL	85.4	86.9	82.4	56.6	91.5	60.5	-	76.5	84.5	74.1	77.6
$\Delta\text{KG-NN}_{uv}$	1.0	0.2	-2.6	3.4	1.2	-4.5	-	-4.8	-0.6	3.9	-0.3
$\Delta\text{KG-NN}_{ut}$	2.4	-1.0	-1.5	-10.1	-32.9	0.2	-	-0.5	-1.6	-1.6	-5.2
$\Delta\text{KG-NN}_{uf}$	1.9	-0.8	-1.3	1.4	-2.4	-0.5	-	3.4	0.9	3.3	0.7
$\Delta\text{KG-NN}_u$	0.4	0.5	-1.9	1.8	-1.5	2.6	-	-1.4	-0.6	2.1	0.2
$\Delta\text{KG-NN}_{sv}$	0.5	-0.9	2.6	-0.6	-0.1	0.5	-	0.5	1.0	0.0	0.4
$\Delta\text{KG-NN}_{st}$	1.0	-2.1	-0.5	1.9	-0.4	0.8	-	0.5	1.6	3.0	0.6
$\Delta\text{KG-NN}_{sf}$	2.7	-0.3	-1.5	-0.7	-1.0	-2.6	-	0.0	0.4	1.8	-0.1
$\Delta\text{KG-NN}_s$	-1.6	-1.0	-2.6	-2.2	-1.2	2.8	-	3.1	1.2	4.3	0.3

Table 6.1: Quantitative evaluation for Cifar10. Comparison of the individual class accuracies for the Cifar10 dataset as training domain and the Stl10 dataset as testing domain. We compare the contextual view-trained DNNs against their baseline SupSSL.

image dataset, and for $\mathbf{h}_{v,s_u^{view}}$ we calculate the class-means of the image representations. Since the goal is to learn contextual image classifiers, we investigate if the context is transferred to $\mathbf{h}_{s^{view}}$ and $\mathbf{h}_{v,s^{view}}$, respectively. It can be seen that semantic relationships provided by the GKG^{view} are reflected in $\mathbf{h}_{s_u^{view}}$. In $\mathbf{h}_{s_u^v}$, the airplane has the highest similarity to the truck and the bird, in $\mathbf{h}_{s_u^t}$, the airplane has the highest similarity to the ship, in $\mathbf{h}_{s_u^f}$, the airplane has the highest similarity to the automobile, and \mathbf{h}_{s_u} the airplane has a high similarity to all vehicles. Further, one notices that taxonomical and generic \mathbf{h}_{s_u} have two main distinctive groups in the embedding space. In $\mathbf{h}_{s_u^t}$ and \mathbf{h}_{s_u} vehicles and animals have a high inter-cluster, but a small intra-cluster variance. For $\mathbf{h}_{v,s_u^{view}}$, we observe that similarities in the GKG^{view} and $\mathbf{h}_{s_u^{view}}$ are only partially reflected. All $\mathbf{h}_{v,s_u^{view}}$ seem to have a similar underlying pattern of the class distribution, with minor differences. We think that implicit relations between class features interfere with the similarities given by \mathbf{h}_{s_u} and the GKG .

Further, we retrieve different distributions for either Cifar10 or Stl10. This behavior can be

explained by the distribution shift between the source and the target domain. While the network attempts to separate classes in the training domain Cifar10, this separation is less successful in the testing domain Stl10.

In Table 6.1 we compare the final object recognition accuracy of the contextual DNNs, compared to their baseline SupSSL. SupSSL is the same model trained with the supervised contrastive loss [250] and without prior context. In Table 6.1a We observe that for different contextual infusions, the overall accuracy is not significantly impacted. For Cifar10 $\Delta\text{KG-NN}_{u^t}$ with -7.0 is the worst performing model, whereas $\Delta\text{KG-NN}_{u^f}$ with 0.4 is the best performing model. We marked the best-performing model for every class in bold. It can be seen that for every class a different contextual model is outperforming the others. It also shows that context influences the focus a DNN puts on predicting a specific class. Table 6.1b shows the relative accuracies of the contextual models on the Stl10 dataset. Note that the models are only trained on Cifar10 data. The goal of that domain generalization scenario is to test the robustness of the models. When evaluated on the target domain, it can be observed that almost in every contextual model the relative accuracy is increased compared to the baseline with no contextual knowledge. In scenarios where the domain changes, we observe strange phenomena occurring such that the model with the second worst performance KG-NN_{u^t} for the class Aircraft of the Cifar10 dataset is the model with the second best performance for Aircraft on Stl10. However, for most of the classes, we see a trend that the best-performing model for a class in Cifar10 tends to perform also better on the target domain.

6.3.3 Experiments on Mini-ImageNet

Dataset Settings: We use Mini-ImageNet, a subset of the ImageNet dataset, as our training domain. It contains 100 classes, each having 600 images of size 84×84 . As the testing domain we use ImageNetV2 [174] comprising 10 new test images per class, ImageNet-Sketch [258] with 50 images per class, ImageNet-R [257], which has 150 images in the style of art, cartoons, deviantart, and ImageNet-A [259] with 7.500 unmodified real-world examples.

Knowledge Graph Construction: Our GKG is built using the three contextual views, namely visual, taxonomical, and functional. GKG^v contains visual properties, e.g. *hasColor*: black, blue, brown; *hasTexture*: dotted, striped, uniform; *hasSize*: large, medium-large, small; and *hasShape*: ellipsoid, quadratic, rectangular. GKG^t contains a taxonomy of the classes using the type-relation. Following DBpedia [149], the class *Malamute* is-a *Dog*, is-a *Mammal*, is-an *Animal*, is-an *Eukaryote*, and is-a *Species*. GKG^f defines the function of a class with properties like: *hasSpeed*: fast, static, slow; *hasWeight*: heavy, light, middle; or *hasTransportation*: goods, none, people. Our GKG contains in total 166 classes, 14 object properties, and 183 individuals.

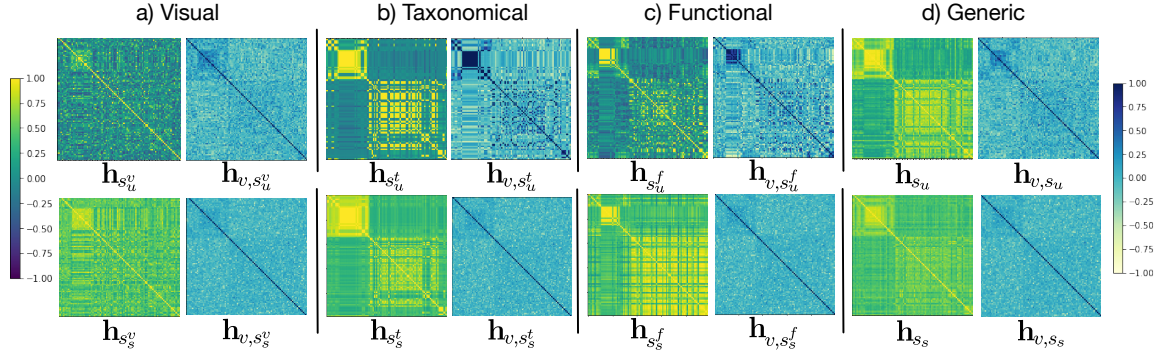


Figure 6.6: Qualitative evaluation for Mini-ImageNet. We compare $\mathbf{h}_{s_u}^{view}$ and $\mathbf{h}_{v,s_u}^{view}$, as well as $\mathbf{h}_{s_s}^{view}$ and $\mathbf{h}_{v,s_s}^{view}$ based on, a) the visual view, b) the taxonomical view, c) the functional view, and d) the full generic KG. To investigate how the semantic relationships are reflected in the embeddings, we illustrate the individual cosine similarities between the classes of the Mini-ImageNet dataset.

Evaluation: Due to the difficulty of deeply investigating 100x100 class similarities, we provide a qualitative overview of the embedding spaces. Figure 6.6 shows a qualitative comparison of $\mathbf{h}_{s^{view}}$ and $\mathbf{h}_{v,s^{view}}$ of a) the visual view; b) the taxonomical view; c) the functional view; and d) the *GKG*. Complementing the experiment in Section 6.3.2, we illustrate the class similarities of \mathbf{h}_{s_s} and \mathbf{h}_{v,s_s} learned using image data as supervision. Interestingly, it can be observed that the similarities in \mathbf{h}_{s_s} and \mathbf{h}_{s_u} follow a similar pattern, but \mathbf{h}_{s_u} seems to have a stronger contrast. However, when investigating the learned image representations in \mathbf{h}_{v,s_s} it is hard to spot the differences between the individual contextual models.

ImageNet	Mini	V2	Sketch	R	A
SupSSL	58.6	43.0	20.3	4.3	1.2
$\Delta\text{KG-NN}_{u^v}$	-0.3	0.0	-0.6	0.2	-0.2
$\Delta\text{KG-NN}_{u^t}$	-19.6	-13.7	-8.8	-2.8	0.0
$\Delta\text{KG-NN}_{u^f}$	-5.2	-3.3	-2.3	-0.7	0.3
$\Delta\text{KG-NN}_u$	0.8	1.6	-0.6	-0.1	-0.1
$\Delta\text{KG-NN}_{s^v}$	0.9	2.3	0.2	0.0	0.3
$\Delta\text{KG-NN}_{s^t}$	1.3	0.6	0.1	0.1	0.0
$\Delta\text{KG-NN}_{s^f}$	0.4	0.4	0.0	-0.1	-0.1
$\Delta\text{KG-NN}_s$	0.5	0.6	0.1	0.0	0.0

Table 6.2: Quantitative evaluation for Mini-ImageNet. Comparison of the contextual view models and their SupSSL baseline on the Mini-ImageNet and its derivatives, *Mini-ImageNet* (Mini), *ImageNetV2* (V2), *ImageNet-Sketch* (Sketch), *ImageNet-R* (R), and *ImageNet-A* (A).

As depicted in Table 6.2, KG-NN_{u^t} and KG-NN_{u^f} are outperformed by the baseline *SupSSL* and the other models with different contextual views by a large margin. In contrast to the Cifar10 experiment where the least performing model is only 8% worse than the baseline, in Mini-ImageNet

the worst is around 34%. Moreover, KG-NN_{s^t} does not suffer from constraints given by GKG^t . This finding confirms the assumption that a joint training softens the constraints of the GKG .



Figure 6.7: Evaluation of sample predictions for Mini-ImageNet. Contextual Predictions of KG-NN_u (GAE) and KG-NN_s (GAT) and their contextual view on Mini-ImageNet. The contextual view influences the image representation and therefore the final prediction for the same input image.

Similar to the example of ambiguous figures, our approach enables DNNs to interpret the same image in various ways using contextual views given by a KG. The results in Figure 6.7 show that for out-of-distribution images the contextual views play a major role in giving reasonable predictions. The idea is that some class confusions are not as critical as others. In that sense, for some tasks it is uncritical to confuse a goose with a house finch as they are both part of the bird family, however confusing a music instrument (oboe), with a dog (malamute) could lead to problems. We also see that KG-NN_u (GAE) and KG-NN_s (GAT) do not necessarily predict the same image given the same context. We believe that further research is needed w.r.t. investigating how to best incorporate context in combination with image data.

6.4 Work Related to Contextual Visual Models based on KG

Contextual information has always been of great interest to improve CV systems. We structure related work into implicit-contextual visual models, explicit-contextual visual models, and contextual KGEs.

Implicit-Contextual Visual Models: Implicit-contextual visual models contextualize relationships between visual features that occur in the image itself. They are used for object priming, where the context defines a prior on the detection parameters [271] or for object detection and segmentation, where boosting is used to relate objects in an image [289]. Wu et al. [290] improved object recognition by processing object regions and context regions in parallel. To overcome the drawback of small receptive fields from standard CNNs, extensions that incorporate visual features from far

image regions [291, 292] or alternative architectures, such as *vision transformers* (ViTs) [293] have been established recently. Moreover, Gao et al. [294] proposed that all modern DNNs are part of the implicit-contextual models since they aggregate contextual information over image regions.

Explicit-Contextual Visual Models: Explicit-contextual visual models use higher-level information like object co-occurrences or semantic concept relationships. They induce additional contextual information that is either not in the dataset or cannot be automatically extracted by the DNN [295]. To create explicit context based on object relations, most methods use scene graphs that describe a scene based on symbolic representations of entities and their spatial and semantic relations. Scene graphs have been applied to the task of collective or group activity recognition [296, 297], object recognition [298, 282], object detection [195, 299] and visual question answering [300]. Label graphs [301] apply fine-grained labels to an image and are used to improve object recognition and reasoning over object relationships [302]. Semantic scene graphs extend scene graphs by textual descriptions and fine-grained labels of a scene [303]. Context-aware zero-shot learning for object recognition [304] or compositional zero-shot learning methods [233] add observed visual primitive states (e.g. old, cute) to objects (e.g. car, dog) to build an embedding space based on visual context. However, scene correlations need to be addressed very carefully, as implicit-contextual models can heavily depend on learned contextual relationships that are only valid for a specific dataset configuration. Therefore, work was already done to decorrelate objects and their visual features to improve model generalization [305].

Contextual Knowledge Graph Embeddings: Whereas our approach extracts the contextual views in a previous step before the actual KGE, there exist works that create contextualized KGEs based on the full KG. Werner et al. [306] introduced a KG embedding over temporal contextualized KG facts. Their recurrent transformer transforms KGEs into contextual embeddings, given the situation-specific factors of the relationship and the subjective history of the entity. Ning et al. [307] proposed a lightweight framework for the usage of context within standard embedding methods. Wang et al. [308] presented a deep contextualized KGE method that learns representations of entities and relations from constructed contextual entity-relation chains. Wang et al. [309] introduced the *contextualized KG embedding method* (CoKE). They propose to take the contextual nature of KGs into account, by learning dynamic, flexible, and fully contextualized entity and relation embeddings.

6.5 Discussion and Insights

With our work, we provided a method to infuse context in form of GKG^{view} into DNNs for visual object recognition. However, knowledge infusion is not straightforward, as inductive biases of ML, such as hyperparameter selection, weight initialization, or dataset dependence, strongly influence the learned representations. To provide an answer for **RQ4** we start by answering the two sub-questions.

RQ4.1: Can context provided in form of a KG influence learning image representations of a DNN, the final accuracy, and the image predictions? - we list the insights obtained from our investigations:

- GKG^{view} **defines class-relationships**. We showed that various contextual views can be extracted from a GKG and that different views lead to different relationships between classes of the dataset.
- $\mathbf{h}_{s^{view}}$ **needs to reflect GKG^{view}** . The embedding method itself also influences the $\mathbf{h}_{KGE^{view}}$ and the performance of the final prediction model. Context can get lost when transferring GKG^{view} into $\mathbf{h}_{s^{view}}$. Hard constraints either in GKG^{view} or produced by the KGE method, e.g. to represent dissimilar classes in $\mathbf{h}_{s^{view}}$ together, can drastically reduce the prediction accuracy.
- $\mathbf{h}_{s_u^{view}}$ **is only partially reflected in $\mathbf{h}_{v,s_u^{view}}$** . Since data-driven approaches have a strong dependence on the dataset distribution, $\mathbf{h}_{s_u^{view}}$ only influences $\mathbf{h}_{v,s_u^{view}}$ to form a hybrid representation. We see that data augmentation weakens the dataset bias and helps to align $\mathbf{h}_{v,s_u^{view}}$ with $\mathbf{h}_{s_u^{view}}$.
- **Joint training reduces the impact of GKG** . Both the learned \mathbf{h}_{v,s_s} and the achieved accuracy values are only slightly affected by the induced GKG . Neither the qualitative evaluation of \mathbf{h}_{v,s_s} nor the quantitative evaluation based on accuracy shows any significant contextual changes.
- **Context shifts the focus on learning specific classes**. We assume that the context constrains the DNN and its hypothesis space. It is known that DNNs tend to memorize spurious correlations that can lead to catastrophic errors in the real world [85]. We think that the task of our contextual models is to prevent exactly these errors. In our experiments, we showed that specific contextual models performed better on specific classes. We assume that context can shift the overall interest of a DNN to predict a certain class.
- **Context rather influences individual image predictions**. Similar to the proposed motivation of how humans interpret ambiguous figures we see context influencing the prediction of difficult or undefinable images in the dataset.

Regarding **RQ4.2:** Can context help to avoid critical errors in domain-changing scenarios where DNNs fail?

- **Context makes more robust against domain changes**. It can be seen that almost every contextual model increases its relative accuracy compared to the baseline when evaluated on the target domain. Moreover, contextual models that performed better on the source dataset tend to perform better if domain change occurs. We argue that GKG^{view} regularizes the strong dependency on the source domain and thus increases the performance on the target domain.

6.6 Summary

In this work, we proposed a framework for context-driven visual object recognition based on KGs. We qualitatively and quantitatively investigated how different contextual views, as well as their

embedding and their infusion method, influence the learned DNN. Further, we have seen that contextual models tend to have a minor impact on the final accuracy, but a major impact on how individual classes or images are represented and predicted. In particular, for out-of-distribution data, where data-driven approaches suffer from less knowledge, contextual image representations help to constrain the hypothesis space, leading to more reasonable predictions. However, there are still challenges to be faced.

We conducted intensive research about a possible context infusion approach and emerging challenges. On the one hand, we have the implementation of the infusion method, which itself heavily depends on modeling choices, weight initialization, as well as network and hyperparameter selection. On the other hand, there is a strong dependence on the image data, which originally comes with an initial dataset bias. This dataset bias limits the ability to influence image data representations and thus predictions influenced by prior knowledge. However, our work showed that with deeper investigations of all the influencing parameters knowledge-infused learning is a promising approach to building context-driven and future intelligent systems.

Chapter 7

Conclusions

The limitations of pure DL approaches have spurred interest in visual transfer learning methods that are less reliant on training data and more adaptable to real-world scenarios. This study explores the incorporation of prior knowledge through KGs to enhance DL’s ability to generalize. KGs offer an effective means of organizing and formalizing knowledge through symbolic operations, making them ideal for knowledge reuse and transfer learning tasks. KGs contain symbolic operations such as logic, rules, and reasoning, and can be created, adapted, and interpreted by human experts.

However, combining the explicit nature of KGs with the implicit vector representations of DL presents various challenges. With KGE methods, that transform the explicit graph representation of a KG into an implicit vector representation of a KGE, new combination possibilities for KG-DL have recently emerged. This work addresses four key challenges in leveraging the prior knowledge encoded in a KG to improve DL for visual transfer learning:

- Enhancing Transfer Learning using Prior Knowledge (**CH1**)
- Investigating Strategies for KG-DL Integration (**CH2**)
- Guiding DL with Prior Knowledge for Visual Transfer Learning (**CH3**)
- Investigating the Impact of Contextual Knowledge for KG-DL (**CH4**)

Overall, the thesis provides a comprehensive investigation into the improvement of DL-based visual transfer learning using KG. The thesis started with Chapter 1, by introducing the problem of visual transfer learning using KG and outlining the four main challenges and research questions. Relevant preliminaries were then presented in Chapter 2, including an overview of NSAI, modalities used in this work, DL-based feature extraction, and inductive biases that affect DNN learning. Then Chapter 3 provides an introduction to transfer learning using prior knowledge and argues that KGs are an ideal representation format for encoding and utilizing prior knowledge. Based on **CH1/RQ1**,

the chapter motivates why a combination of KG and DL is well suited for improved DL and visual transfer learning systems. Chapter 4 focuses on **CH2/RQ2**, by providing a comprehensive categorization of KG-DL combinations. Therefore, it summarizes CV models that already utilize KGs with DL and proposes a categorization based on their knowledge integration. Chapter 5 investigates how prior knowledge encoded in a KG can guide DL to improve visual transfer learning, based on **CH3/RQ3**. In particular, it provides a novel KG-DL method for visual transfer learning of the *KG as a Trainer* category. Chapter 6 focuses on **CH4/RQ4**, by investigating the impact of different types of prior knowledge encoded in a KG on the performance of KG-DL, especially for visual transfer learning. Finally, Chapter 7 concludes the thesis, by revisiting all research questions, posing further challenges and open issues, and providing an outlook for future work. Overall, the thesis contributes to the understanding of how prior knowledge encoded in KGs can be used to improve DL-based visual transfer learning and provides insights into the challenges and opportunities of such an integration.

7.1 Revisiting the Research Questions

In this thesis, we have intensively studied the problem of visual transfer learning using KG. Within this section, we will therefore provide comprehensive answers to each of the research questions formulated in Section 1.3:

RQ1: Why can prior knowledge encoded in a KG improve DL-based visual transfer learning?

Transfer learning is a powerful approach to learning that enables the application of knowledge gained from one domain to another but related domain. In Chapter 3, we explored the fundamental principles of transfer learning and showed that any DL task relies on transfer learning by design. DL learns features from constrained training data and applies them in reality. However, the information of the training domain is constrained and will always differ from the information in the real world. To generalize gathered knowledge of the training data to unknown domains, DL requires additional prior assumptions about the world [4]. These prior assumptions, known as inductive biases, are pre-assigned and non-learnable components of the model that influence which features are learned from the data. Inductive biases can be considered a form of prior knowledge, however, they are abstract, difficult to interpret, and cannot be easily combined with human expert knowledge.

In general, prior knowledge refers to any knowledge that exists about the problem domain before the learning process begins and can occur in explicit and implicit forms. We showed that a KG can represent prior knowledge in both forms, explicitly and implicitly. KGs naturally work with explicit symbolic knowledge that is encoded in a graph such that humans can understand and interact with it. KGs are flexible, interpretable, and modifiable, and come with an established toolset to

build, reason, interact, and embed heterogeneous sources of information. They unify human expert knowledge with established knowledge from taxonomies, books, and other sources of information and encode it by modeling classes, entities, and their relationships of a domain. With the recent development of KGE methods that transform a KG into vector space, its prior knowledge can be transformed into an implicit representation format. This makes a KG an ideal representation format for prior knowledge, as it interfaces with explicit human knowledge through the KG and connects to implicit embeddings of data from other modalities, such as vision or language, through its KGE.

By leveraging prior knowledge encoded in a KG, DL will learn more efficiently and robustly for visual transfer learning scenarios that naturally occur in the real world. The incorporation of prior knowledge in a KG can help to overcome the limitations of constrained training data and enables the transfer of knowledge across related visual domains, resulting in better generalization and improved performance in real-world applications.

RQ2: What are possible ways of integrating the prior knowledge encoded in a KG into the DL pipeline?

In Chapter 4, we explored different ways of integrating prior knowledge encoded in a KG into the DL pipeline for visual transfer learning. These approaches differ in terms of how they integrate prior knowledge, and we identified four main categories of approaches that combine a KG with DL.

The first category is called *Knowledge Graph as a Reviewer*, in which the KG is used for post-validation. This involves matching the predictions of a DNN with the structured knowledge in the KG to identify any inconsistencies or errors in the output of the visual model. The second category is *Knowledge Graph as a Trainee*, which learns a KGE method that embeds the KG into a KGE $\mathbf{h}_{s,v}$ based on the visual embedding \mathbf{h}_v . Approaches of this category use the visual embedding of the DNN as a guide for training the $\mathbf{h}_{s,v}$ and force the KGE method to associate the graph structure of the KG with the visual embedding of the DNN. Approaches of this category are mostly used for transfer learning scenarios with an output domain change, such as zero-shot or few-shot learning. The third category is *Knowledge Graph as a Trainer*, which uses the KG with its KGE \mathbf{h}_s to learn a DNN that embeds images in a visual-semantic embedding $\mathbf{h}_{v,s}$ that is based on the KGE. This approach uses the structured prior knowledge in the KG to train a DNN that learns its visual embedding based on the representation of the KGE. Approaches within this category can use prior knowledge to influence how a DNN learns from unstructured image data and thus can be used for transfer learning scenarios where input domain change occurs. The fourth category is the *Knowledge Graph as a Peer*, which combines the KG and its KGE \mathbf{h}_s with the visual embedding \mathbf{h}_v to form a hybrid embedding \mathbf{h}_h . In this approach, both the KGE method and the DNN are learned using a joint optimization process. However, at the same time, the approach loses the hard constraints of each modality and its interpretability.

In each of the KG-DL categories, we distinguish between a feature extractor and a transformation model. When an approach retrains the entire encoder network using the other modality, we refer to it as a feature extractor. When it learns only a transformation function from the embedding of one modality to another, we define it as a transformation model. All categories provide different ways of integrating prior knowledge encoded in a KG into the DL pipeline and can be used differently to improve the performance of visual transfer learning.

RQ3: How can prior knowledge encoded in a KG guide DL to improve visual transfer learning?

In Chapter 5, we explored the use of prior knowledge encoded in a KG to guide DL to improve visual transfer learning. We proposed the KG-NN approach, which belongs to the category *KG as a Trainer*. This approach leverages prior knowledge of a KG to influence the learning of certain features from image data in a DNN. By doing so, it can improve generalization and efficiency in transfer learning scenarios.

The KG-NN approach works by constructing a KG based on prior knowledge about the domain, including individual properties and relationships between classes. The KG is then converted into a KGE using an appropriate KGE method to align it with the implicit knowledge of the visual embedding from the DNN. The KG-NN approach optimizes the entire feature extractor of the DNN by attracting the visual embedding of the images to their representatives in the KGE using a contrastive loss function. By doing so, it forces the DNN to learn visual features that depend on the relationships in the KGE and are thus partially independent of the training data distribution.

We compared KG-NN with approaches without prior knowledge and demonstrated that it outperforms them, particularly in transfer learning tasks. KG-NN not only provides better generalization but also learns more efficiently, requiring significantly less training data to adapt to a new domain. We tested the approach in various domains, such as road sign recognition, where the KG defines the relationships between classes based on prior knowledge controlled by human experts. The KG-NN approach can be applied to any other domains where prior knowledge is available that can be encoded in a KG. Overall, KG-NN demonstrates the potential of using prior knowledge encoded in a KG to guide DL for improving visual transfer learning.

RQ4: How does the type of prior knowledge in a KG impact KG-DL performance, especially in visual transfer learning?

We have shown that prior knowledge in KG-DL can help improve visual transfer learning by learning more generalizable representations from unstructured image data. However, the question

remains how different types of prior knowledge affect the behavior of KG-DL and its results. Therefore, Chapter 6 investigated the impact of contextual prior knowledge on a DNN. Similar to the human brain, which encodes visual inputs and their objects into contextual embeddings, we constructed experiments to learn different DNNs based on the same visual input but using different contextual knowledge of a KG. Whereas objects in the brain can be embedded based on their visual properties, tools are encoded based on their functional properties. While pure DL can only encode images based on their visual properties, KG-DL can induce any contextual knowledge using the KG.

Therefore, we provide three different types of context, i.e. contextual views, namely visual, taxonomic, and functional context. The induced contextual view influences the neighborhood relationships between images and classes in the embedding space and we see that the contextual KG-DL models perform differently based on their induced context. It can be seen that the overall accuracy, but also the accuracy of individual classes and the prediction for specific image samples changes. In addition, hard constraints of the KG that contradict the information in the images can lead to optimization problems and problems in distinguishing between specific classes. However, we also see that joint optimization methods, such as those provided by *Knowledge Graph as a Peer* methods, relax these hard constraints. Despite this, adding any contextual prior knowledge improves generalizability and thus the performance on visual transfer learning tasks. However, there are many factors besides contextual views that influence the learning of a DNN, such as the encoding format of prior knowledge in the KG, as well as the KGE method that converts the graph-based KG into a vector-based format. Moreover, DL is always highly influenced by various inductive biases, such as the dataset itself, the augmentation method, the labels, the DNN architecture, the loss function, and the optimization method, which make analyzing the exact influence of contextual prior knowledge and fully controlling the learning of a DNN a difficult task.

In summary, we have shown that KG-DL performance depends on contextual view, the infusion method, the inductive biases, and the final task. The type of prior knowledge, i.e. the contextual view, affects the predictions for individual classes and images, and the importance of context increases if the domain changes.

7.2 Further Challenges and Open Issues

In this work, we have shown the potential of combining KG with DL, especially for visual transfer learning tasks. We have seen that adding prior knowledge to data-driven DL provides a way to influence the learning process and helps DNNs to learn more generalizable models. For this purpose, we worked on several challenges for KG-DL. We argued why a KG can improve DL-based visual transfer learning, introduced different categories for knowledge integration in KG-DL, developed a specific KG-DL method for visual transfer learning, and studied the impact of contextual prior knowledge of a KG on the performance of KG-DL. However, there are still many open challenges

for visual transfer learning using KG that need to be further investigated.

Relevant Knowledge and its Representation: As explained in Section 2.2.3, a KG can be created using different modeling structures, e.g., directed, labeled, hyper- or hyper-relational graphs. It is important to investigate how these structures affect the KGE and the final performance of KG-DL. Thus, the open challenge is to find a balance between the level of detail of the relevant knowledge and the complexity of the structures used to represent it.

In addition, as highlighted in Chapter 6, different types of prior knowledge lead to different behavior and results of KG-DL, but cannot be fully controlled. Therefore, the relationship between specific prior knowledge and DL behavior needs to be explored in more detail. We anticipate that it will be necessary to include the DL pipeline with all its inductive biases, and the task in the process of selecting relevant knowledge. Therefore, a generic KG that contains multiple task-dependent contextual views must be able to quickly retrieve relevant compositions of knowledge.

In summary, it is critical to understand how KG modeling structures influence KGE and consequently KG-DL performance, but also how the task and DL pipeline itself influence what prior knowledge is relevant.

Knowledge Graph Embedding and its Metrics: In the context of this thesis, we have investigated various KGE methods, with the objective of understanding how the KGE method impacts the KGE, the induction of prior knowledge, and the final KG-DL performance.

In Section 2.3.2, we have seen that the choice of a KGE method is highly dependent on how the knowledge is modeled in the KG. For instance, most GNN-based KGE methods expect directed labeled graphs as input and have difficulties in encoding n-ary relations, edge attributes, literals, etc. Moreover, it needs to be investigated how KGE methods can work with multimodality and how hierarchies, rules, and concepts from a KG can be maintained after embedding the KG into KGE.

In addition, typical metrics for KGE, such as AMR, MRR, Hits@K, do not reflect the final performance of KG-DL when combined with image data. Therefore, it is important to investigate which type of KGE method is suitable for which task and develop metrics to decide if a KGE is suitable for knowledge induction into a visual model.

Transforming the KG into KGE using joint training with image data representations of the DNN, similar to the category *KG as a Peer*, seems to be another direction, but also lowers the impact of the KGE. The challenge here is to find a perfect balance between supporting the optimization of the KGE with image data and constraining the learning process of the DNN.

Impact of Prior Knowledge and Inductive Biases: Data-driven DL approaches derive their implicit knowledge from the training data only. KG-NN brings additional prior knowledge about KG into the learning process by using a training objective that aligns the two embedding spaces. However, recent studies [94] showed that the objective function, and so the prior knowledge, mainly

influences the last layers of a DNN. This means that the learned representations of the early network layers are only partially influenced by prior knowledge. It needs to be investigated if this limits the performance of KG-NN and if alternative infusion techniques can bring benefit.

With the goal of fully controlling the DNN using prior knowledge encoded in a KG, we have seen in Chapter 6 that the behavior of DNNs is also highly dependent on the choice of its inductive biases. Therefore, the prior knowledge induced by the KG is only one piece of the puzzle and must be reconciled with all other training effects. To further analyze DNN's behavior when inducing prior knowledge of a KG, ways to consider inductive biases need to be found. In summary, the specific influence of prior knowledge on DNN layers needs to be further investigated, and alternative induction strategies or extensions to KG-NN that take inductive biases into consideration need to be explored.

7.3 Future Work

The combination of KG with DL is based on the discussion of combining symbolic AI with ML, and the deeper-rooted question of how to achieve artificial human intelligence. We hope that this thesis will help the reader to systematically combine the technology of KG with DL for developing models that benefit from the appropriate combination of semantic prior knowledge with visual information. In this work, we have shown that a KG is able to influence the behavior of a DNN. If we further learn how to fully control how the DNN learns from unstructured data, there is an opportunity to overcome the drawbacks of DL and improve future AI.

To achieve the goal of effectively using KG with DL, it will be important to further investigate the impact of knowledge structures, the relevance of KGE methods, and the role of inductive biases in the learning process. If KG-DL is supposed to be used for better and safer products, extensions to our method need to be developed that leverage data at scale, include sample-based in addition to the class-based context in the KG, fully constrain the DNN, induce explicit rules, use temporal knowledge, deal with evolving knowledge, and leverage the KG in combination with the DNN at inference time.

Scaling Models: Humans have been shown to be significantly intellectually superior to apes, yet cognitive scientists assume that the human brain may just be a scaled-up version of the primate brain [310]. The fact that scale is essential for the development of intelligent systems is also evident in recent applications of AI, such as foundation models. Foundation models are huge DNNs trained on large and diverse datasets, outperforming more specialized and optimized, but also smaller, models in most downstream NLP and CV tasks [122]. However, obtaining labeled training data can be a rather tedious, error-prone, and sometimes impossible process.

Therefore, foundation models use various tricks to train without the need for additional expensive

labels, enabling them to leverage information from vast amounts of data. Foundation models for CV, e.g. CLIP [86], use the fact that most images on the Internet co-occur with additional text descriptions. CLIP uses these noisy text annotations as labels, assuming that the content of the image is at least partially described by the noisy text annotation. To relate all the heterogeneous text descriptions based on their underlying meaning, an LLM is used to transform the text description into a lower-dimensional semantic embedding. Similar to the category *KG as a Trainer*, CLIP then uses the semantic embedding to train the visual DNN. Since the major difference between CLIP and KG-NN, is the availability of large amounts of free image-text pairs, one future direction is to explore where prior KG-image pairs are already available or how they could be collected automatically.

In particular, areas such as autonomous driving, the production line, or the *Internet of things* (IoT) are promising, since huge amounts of graph-based metadata could be automatically collected with images. Moreover, we assume that a foundation model of KG-DL would lead to even more generalized features and higher accuracy, especially in visual transfer learning tasks.

Unifying Sample-based and Class-based Context: In the scope of this work, KG-NN is only trained with class-based context, i.e. each image of a class has the same representation in KGE due to its properties and its relationship to other classes in the KG. More specifically, the KG contains class-based context due to higher-level class properties, e.g., *stop sign - has shape - octagon*, *stop sign - has background color - red*, etc. However, class-based context needs expensive data labeling to define the relevant class for the images and the assumption that the information in the image is just defined by the class does not fully reflect the real world.

In reality, images of different classes can also be similar due to other visual properties. We refer to such visual properties of images as sample-based context, that describe the properties of an individual image and therefore its relationship to other images, e.g., *has camera*, *has environment*, *has location*, etc.

Moreover, Sample-based context could be manually annotated or automatically collected in many real-world applications of CV. The KG can include higher-level concepts, relating the class-based context with the sample-based context, to provide further meaning to noisy collected data. When using the *KG as a Trainer*, each image will be assigned to a specific contextual position in the semantic embedding space based on its sample and class-based context. Furthermore, we argue that unifying class-based context and sample-based context in a KG will help KG-DL models to further improve accuracy and generalization.

Constraining Models and Inducing Explicit Rules: If models should be deployed into safety-relevant products, such as autonomous driving, they should be able to follow certain constraints and obey certain rules. One of the main goals of NSAI is to enable DL to operate on fixed structures where constraints, rules, and reasoning can be applied. We argue that inducing such a structure, e.g. an ontology or a KG, into the learning process will enable DL to learn more efficiently and

constrained from data. Rules, as abstractions derived from general experience, generalize well to related domains. KG-NN uses the KG to induce additional prior knowledge into DL-based learning from unstructured data. However, when transferring the KG to a vector-based KGE, the explicit rules and constraints of the KG are omitted and, if present, are only implicit.

Therefore, methods should be explored that can constrain the huge hypothesis space using structured and robust KGs in combination with unstructured text and images. Methods that can learn implicit representations from data based on explicit rules and semantic structures, will pave the way for future machine intelligence.

Using Temporal Knowledge: Temporal data provides valuable information about the motion and acceleration of objects, essential for object detection and tracking. However, temporal data also provide important information about the relationship between images, objects, and features that cannot be described without temporal dependency. Shuffling data and pretending that the image frames are i.i.d. is a false assumption and a loss of information [81]. Therefore, acquiring the temporal information is important to identify relevant correlations and avoid learning spurious correlations from the data [85]. However, video data is very complex, and DL methods can only effectively analyze single frames due to computational limitations.

We argue that KGs can serve as temporal memories that store contextual information about images, such as metadata features or objects, across multiple timestamps. If KG-DL would be able to reuse the information from a temporal KG memory during training, important correlations in single image frames could be identified, even after shuffling the data. In a temporal KG-DL, metadata for the images can be transferred to a KG that links and stores temporal information. When training the DNN, this KG metadata can be reused to allow the DNN to extract temporally relevant features and avoid learning spurious correlations.

In summary, such a method combining a temporal KG and DL can leverage temporal information to distinguish between relevant and spurious features in image data. Having such a system would enable DL to be more accurate and more robust when predicting in the real world.

Evolving Knowledge: Knowledge of the real world or specific knowledge of a domain can evolve rapidly. The main problem of DNNs is that they require a time-consuming and resource-intensive re-training process when updated with new data. For KG-NN, where the KG influences how a DNN learns, the same problem applies.

However, there are ideas on how KG-DL can deal with evolving knowledge. One possibility to deal with evolving knowledge is to extract implicit knowledge from the DNN [17] or heterogeneous knowledge from external sources and integrate it into a KG that can be updated, managed, and refined by humans. Progress has already been made in the area of KG construction by embedding and information extraction methods [311, 312, 313]. This would allow the application of rules and reusable knowledge structures over detected visual features in the images.

If the goal is to process evolving unstructured data, methods must be developed to extract implicit features from DNNs, integrate these features into KGs, and either induce the evolved knowledge of the KG back into the DNN during inference or use the knowledge of the KG directly for prediction.

Using the Knowledge Graph at Inference: DNNs struggle or behave unpredictably when presented with noisy visual data during inference. In contrast, humans are able to make robust and accurate predictions by leveraging multiple sources of information at inference. For example, when driving at night on a highway, humans can predict the presence of a car by simply seeing two lights in front of them, using commonsense to fill in the gaps in their visual input.

However, DL systems rely solely on visual input and therefore fail if that input is noisy or degraded. To address this challenge, it is important to explore ways to incorporate additional knowledge into the system at inference time, especially in situations where training data is scarce.

We propose three different ideas to leverage a KG to improve the performance of DNNs during inference. The first idea involves using the *KG as a Reviewer*, where the plausibility of the output of the DNN is evaluated and constrained by the KG. The second idea involves model-agnostic fusion of modalities, in which the DNN stacks image data with vector-based prior knowledge of the KG and automatically learns how best to combine the two to solve the task. Finally, the third idea involves infusing the prior knowledge directly into the embedding space, allowing the two modalities to be combined in a way that optimizes the performance of the DNN. However, whether these methods will be helpful to improve KG-DL performance at inference needs further investigation.

Although incorporating prior knowledge from a KG into a DL system at inference is not straightforward, it will be essential to improve KG-DL performance in situations where visual data is noisy or scarce, and where humans are able to make accurate predictions by leveraging multiple sources of information.

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und nur mit den angegebenen Hilfsmitteln verfasst habe und die wörtlich oder dem Inhalt nach aus fremden Arbeiten entnommenen Stellen als solche kenntlich gemacht sind. Ferner versichere ich, dass ich die gleiche Arbeit noch nicht für eine andere wissenschaftliche Prüfung eingereicht und mit der gleichen Abhandlung weder bereits einen Doktorgrad erworben noch einen Doktorgrad zu erwerben versucht habe.

(Sebastian Monka)

Bibliography

- [1] OED Online. *artificial intelligence, n.* Oxford University Press, 2022.
- [2] Allen Newell and Herbert A. Simon. Computer science as empirical inquiry: Symbols and search. *Commun. ACM*, 19(3):113–126, mar 1976.
- [3] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *CoRR*, abs/1604.00289, 2016.
- [4] Max Welling. Do we still need models or just more data and compute? *University of Amsterdam*, 2019.
- [5] James J. DiCarlo and David D. Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 2007.
- [6] Michelle R. Greene and Bruce C. Hansen. Disentangling the independent contributions of visual and conceptual features to the spatiotemporal dynamics of scene categorization. *bioRxiv*, 2020.
- [7] Susan G. Wardle and Chris Ian Baker. Recent advances in understanding object recognition in the human brain: deep neural networks, temporal dynamics, and context. *F1000Research*, 9, 2020.
- [8] François Chollet. On the measure of intelligence. *CoRR*, abs/1911.01547, 2019.
- [9] Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 2011.
- [10] Grady Booch, Francesco Fabiano, Lior Horesh, Kiran Kate, Jonathan Lenchner, Nick Linck, Andrea Loreggia, Keerthiram Murugesan, Nicholas Mattei, Francesca Rossi, and Biplav Srivastava. Thinking fast and slow in AI. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 15042–15046. AAAI Press, 2021.

- [11] Artur S. d'Avila Garcez and Luís C. Lamb. Neurosymbolic AI: the 3rd wave. *CoRR*, abs/2012.05876, 2020.
- [12] Nicolas Palanca-Castan, Beatriz Sánchez Tajadura, and Rodrigo Cofré. Towards an interdisciplinary framework about intelligence. *Heliyon*, 7(2):e06268, 2021.
- [13] Plato. Diálogos iv. *Republica (Republic)*, Gredos, 1988.
- [14] Plato. Phaedrus. *Gredos*, 2011.
- [15] Aristotele. De anima. 350 BC.
- [16] Henry A. Kautz. The third AI summer: AAAI robert s. engelmore memorial lecture. *AI Mag.*, 43(1):93–104, 2022.
- [17] Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.
- [18] Raymond B Cattell. Raymond b. cattell. In *A history of psychology in autobiography, Vol VI*, pages 61–100. Prentice-Hall, Inc, Englewood Cliffs, 2007.
- [19] K S Mcgrew. The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D P L Flanagan & P, editor, *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, pages 136–181. The Guilford Press, 2005.
- [20] Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, January 2007.
- [21] A.L. Hodgkin and A.F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117:500–544, 1952.
- [22] Gary Marcus. Deep learning: A critical appraisal. *CoRR*, abs/1801.00631, 2018.
- [23] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, December 1943.
- [24] G L Shaw. Donald hebb: The organization of behavior. In *Brain Theory*, pages 231–233. Springer Berlin Heidelberg, 1986.
- [25] F Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.*, 65(6):386–408, November 1958.
- [26] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986.

- [27] Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, 2006.
- [28] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 153–160. MIT Press, 2006.
- [29] Marc’Aurelio Ranzato, Christopher S. Poultney, Sumit Chopra, and Yann LeCun. Efficient learning of sparse representations with an energy-based model. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 1137–1144. MIT Press, 2006.
- [30] Gary Marcus and Ernest Davis. Insights for AI from the human mind. *Commun. ACM*, 64(1):38–41, 2021.
- [31] Geoffrey E. Hinton. How to represent part-whole hierarchies in a neural network. *Neural Comput.*, 35(3):413–452, 2023.
- [32] Chaim Zins. Conceptual approaches for defining data, information, and knowledge. *J. Assoc. Inf. Sci. Technol.*, 58(4):479–493, 2007.
- [33] Rafael C. Gonzalez and Richard E. Woods. *Digital image processing*. Prentice Hall, Upper Saddle River, N.J., 2008.
- [34] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. *Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges*. 2021.
- [35] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.
- [36] Adrian Akmajian. *Linguistics - An Introduction to Language and Communication*. MIT Press, Cambridge, 2010.
- [37] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler,

- Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
- [38] Egon Werlich. A text grammar of english. 1976.
- [39] Yann LeCun. A Path Towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27.
- [40] John F. Sowa. Semantic Networks. 1987.
- [41] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *CoRR*, abs/2003.02320, 2020.
- [42] Renzo Angles and Claudio Gutierrez. An introduction to graph data management. In George H. L. Fletcher, Jan Hidders, and Josep-Lluís Larriba-Pey, editors, *Graph Data Management, Fundamental Issues and Recent Developments*, Data-Centric Systems and Applications, pages 1–32. Springer, 2018.
- [43] Renzo Angles, Harsh Thakkar, and Dominik Tomaszuk. Mapping RDF databases to property graph databases. *IEEE Access*, 8:86091–86110, 2020.
- [44] Christopher M. Bishop. *Pattern recognition and machine learning, 5th Edition*. Information science and statistics. Springer, 2007.
- [45] Xuechuan Wang and Kuldip K. Paliwal. Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. *Pattern Recognit.*, 36(10):2429–2439, 2003.
- [46] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.
- [47] Koray Kavukcuoglu, Marc'Aurelio Ranzato, Rob Fergus, and Yann LeCun. Learning invariant features through topographic filter maps. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 1605–1612. IEEE Computer Society, 2009.
- [48] Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016.

- [49] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020.
- [50] Trevor Hastie, Jerome H. Friedman, and Robert Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2001.
- [51] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.
- [52] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 886–893. IEEE Computer Society, 2005.
- [53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.
- [54] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Trans. Knowl. Data Eng.*, 29(12):2724–2743, 2017.
- [55] Ines Chami, Sami Abu-El-Haija, Bryan Perozzi, Christopher Ré, and Kevin Murphy. Machine learning on graphs: A model and comprehensive taxonomy. *CoRR*, abs/2005.03675, 2020.
- [56] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. A survey on knowledge graphs: Representation, acquisition and applications. *CoRR*, abs/2002.00388, 2020.
- [57] M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734 vol. 2, 2005.
- [58] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.
- [59] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

- [60] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4869–4880, 2019.
- [61] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795, 2013.
- [62] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.
- [63] Mikhail Galkin, Etienne G. Denis, Jiapeng Wu, and William L. Hamilton. Nodepiece: Compositional and parameter-efficient representations of large knowledge graphs. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.
- [64] Paolo Rosso, Dingqi Yang, and Philippe Cudré-Mauroux. Beyond triplets: Hyper-relational knowledge graph embedding for link prediction. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen, editors, *WWW ’20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 1885–1896. ACM / IW3C2, 2020.
- [65] Bahare Fatemi, Perouz Taslakian, David Vázquez, and David Poole. Knowledge hypergraphs: Prediction beyond binary relations. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 2191–2197. ijcai.org, 2020.
- [66] Jianfeng Wen, Jianxin Li, Yongyi Mao, Shini Chen, and Richong Zhang. On the representation and embedding of knowledge bases beyond binary relations. In Subbarao Kambhampati, editor, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 1300–1307. IJCAI/AAAI Press, 2016.
- [67] Huan Gui, Jialu Liu, Fangbo Tao, Meng Jiang, Brandon Norick, and Jiawei Han. Large-scale embedding learning in heterogeneous event data. In Francesco Bonchi, Josep Domingo-Ferrer, Ricardo Baeza-Yates, Zhi-Hua Zhou, and Xindong Wu, editors, *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, pages 907–912. IEEE Computer Society, 2016.

- [68] Chia-An Yu, Ching-Lun Tai, Tak-Shing Chan, and Yi-Hsuan Yang. Modeling multi-way relations with hypergraph embedding. In Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang, editors, *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 1707–1710. ACM, 2018.
- [69] Jie Huang, Chuan Chen, Fanghua Ye, Jiajing Wu, Zibin Zheng, and Guohui Ling. Hyper2vec: Biased random walk for hyper-network embedding. In Guoliang Li, Jun Yang, João Gama, Juggapong Natwichai, and Yongxin Tong, editors, *Database Systems for Advanced Applications - 24th International Conference, DASFAA 2019, Chiang Mai, Thailand, April 22-25, 2019, Proceedings, Part III, and DASFAA 2019 International Workshops: BDMS, BDQM, and GDMA, Chiang Mai, Thailand, April 22-25, 2019, Proceedings*, volume 11448 of *Lecture Notes in Computer Science*, pages 273–277. Springer, 2019.
- [70] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3558–3565. AAAI Press, 2019.
- [71] Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha P. Talukdar. Hypergcn: A new method for training graph convolutional networks on hypergraphs. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1509–1520, 2019.
- [72] Ke Tu, Peng Cui, Xiao Wang, Fei Wang, and Wenwu Zhu. Structural deep embedding for hyper-networks. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 426–433. AAAI Press, 2018.
- [73] Inci M. Baytas, Cao Xiao, Fei Wang, Anil K. Jain, and Jiayu Zhou. Heterogeneous hyper-network embedding. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*, pages 875–880. IEEE Computer Society, 2018.

- [74] Ruochi Zhang, Yuesong Zou, and Jian Ma. Hyper-sagmn: a self-attention based graph neural network for hypergraphs. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- [75] Richong Zhang, Junpeng Li, Jiajie Mei, and Yongyi Mao. Scalable instance reconstruction in knowledge bases via relatedness affiliated embedding. In Pierre-Antoine Champin, Fabien L. Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis, editors, *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1185–1194. ACM, 2018.
- [76] Yu Liu, Quanming Yao, and Yong Li. Generalizing tensor decomposition for n-ary relational knowledge bases. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen, editors, *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 1104–1114. ACM / IW3C2, 2020.
- [77] Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. Tucker: Tensor factorization for knowledge graph completion. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5184–5193. Association for Computational Linguistics, 2019.
- [78] Saiping Guan, Xiaolong Jin, Yuanzhuo Wang, and Xueqi Cheng. Link prediction on n-ary relational data. In Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 583–593. ACM, 2019.
- [79] Mikhail Galkin, Priyansh Trivedi, Gaurav Maheshwari, Ricardo Usbeck, and Jens Lehmann. Message passing for hyper-relational knowledge graphs. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7346–7359. Association for Computational Linguistics, 2020.
- [80] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha P. Talukdar. Composition-based multi-relational graph convolutional networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- [81] Leon Bottou. Learning Representation with Causal Invariance (ICLR 2019). <https://leon.bottou.org/talks/invariances>.

- [82] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, April 1997.
- [83] Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *CoRR*, abs/2011.15091, 2020.
- [84] Laura von Rügen, Sebastian Mayer, Jochen Garcke, Christian Bauckhage, and Jannis Schücker. Informed machine learning - towards a taxonomy of explicit integration of knowledge into machine learning. *CoRR*, abs/1903.12394, 2019.
- [85] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. Technical report, arXiv:1907.02893, 2019.
- [86] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *Image*, 2:T2, 2021.
- [87] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.
- [88] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60, 2019.
- [89] Randall Balestriero, Léon Bottou, and Yann LeCun. The effects of regularization and data augmentation are class dependent. *CoRR*, abs/2204.03632, 2022.
- [90] Kaize Ding, Zhe Xu, Hanghang Tong, and Huan Liu. Data augmentation for deep graph learning: A survey. *CoRR*, abs/2202.08235, 2022.
- [91] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, 2015.
- [92] Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. A unifying mutual information view of metric learning: Cross-entropy vs. pairwise losses. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, pages 548–564. Springer, 2020.
- [93] Simon Kornblith, Honglak Lee, Ting Chen, and Mohammad Norouzi. What’s in a loss function for image classification? *CoRR*, abs/2010.16402, 2020.

- [94] Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 28648–28662, 2021.
- [95] John S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In Françoise Fogelman-Soulié and Jeanny Héroult, editors, *Neurocomputing - Algorithms, Architectures and Applications, Proceedings of the NATO Advanced Research Workshop on Neurocomputing Algorithms, Architectures and Applications, Les Arcs, France, February 27 - March 3, 1989*, volume 68 of *NATO ASI Series*, pages 227–236. Springer, 1989.
- [96] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 935–943, 2013.
- [97] L. V. D. Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
- [98] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 1735–1742. IEEE Computer Society, 2006.
- [99] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomáš Mikolov. Devise: A deep visual-semantic embedding model. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2121–2129, 2013.
- [100] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1386–1393. IEEE Computer Society, 2014.
- [101] Herbert E. Robbins. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

- [102] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [103] D. Rumelhart, Geoffrey E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.
- [104] Geoffrey E. Hinton. The forward-forward algorithm: Some preliminary investigations. *CoRR*, abs/2212.13345, 2022.
- [105] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for boltzmann machines. *Cogn. Sci.*, 9(1):147–169, 1985.
- [106] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13:281–305, 2012.
- [107] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and D. Mike Titterton, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pages 249–256. JMLR.org, 2010.
- [108] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1026–1034. IEEE Computer Society, 2015.
- [109] Alexander D’Amour, Katherine A. Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, and Jonathan Deaton. Underspecification presents challenges for credibility in modern machine learning. *CoRR*, 2020.
- [110] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *ICANN*, 2018.
- [111] David Perkins and Gavriel Salomon. Transfer of learning. 11, 07 1999.
- [112] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proc. IEEE*, 109(1):43–76, 2021.
- [113] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, 2010.
- [114] Sebastian Ruder and Barbara Plank. Learning to select data for transfer learning with bayesian optimization. In *EMNLP*, 2017.

- [115] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 2178–2186, 2011.
- [116] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 10–18. JMLR.org, 2013.
- [117] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In Vera Kurková, Yannis Manolopoulos, Barbara Hammer, Lazaros S. Iliadis, and Ilias Maglogiannis, editors, *Artificial Neural Networks and Machine Learning - ICANN 2018 - 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III*, volume 11141 of *Lecture Notes in Computer Science*, pages 270–279. Springer, 2018.
- [118] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [119] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.
- [120] Alexander D’Amour, Katherine A. Heller, Dan Moldovan, Ben Adlam, and et. al. Underspecification presents challenges for credibility in modern machine learning. *CoRR*, abs/2011.03395, 2020.
- [121] Jianyu Zhang, David Lopez-Paz, and Léon Bottou. Rich feature construction for the optimization-generalization dilemma. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 26397–26411. PMLR, 2022.
- [122] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, and et al. On the Opportunities and Risks of Foundation Models, July 2022.

- [123] Gregory Murphy and E.J. Winiewski. Feature correlations in conceptual representations. In G Tiberghien, editor, *Advances in Cognitive Science*, volume 2, pages 23–45. Ellis Horwood, Chichester, 1989.
- [124] A. S. Kaplan and G. L. Murphy. Category learning with minimal prior knowledge. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 26(4):829–846, July 2000.
- [125] Joseph J. Williams and Tania Lombrozo. Explanation and prior knowledge interact to guide learning. *Cognitive Psychology*, 66(1):55–84, February 2013.
- [126] Anmei Dong, Morris Siu-Yung Jong, and Ronnel B. King. How Does Prior Knowledge Influence Learning Engagement? The Mediating Roles of Cognitive Load and Help-Seeking. *Frontiers in Psychology*, 11, 2020.
- [127] Edward S. Shapiro. *Academic Skills Problems: Direct Assessment and Intervention, 4th Ed.* Academic Skills Problems: Direct Assessment and Intervention, 4th Ed. Guilford Press, New York, NY, US, 2011.
- [128] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In Dieter Fox and Carla P. Gomes, editors, *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, pages 646–651. AAAI Press, 2008.
- [129] Michael Polanyi. Personal Knowledge: Towards a Post-Critical Philosophy.
- [130] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [131] Marcus Rohrbach, Sandra Ebert, and Bernt Schiele. Transfer learning in a transductive setting. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 46–54, 2013.
- [132] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pages 48–64. Springer, 2014.
- [133] Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. In *2018 IEEE Conference on Computer Vision*

- and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1576–1585. IEEE Computer Society, 2018.
- [134] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6857–6866. IEEE Computer Society, 2018.
- [135] Yuxia Geng, Jiaoyan Chen, Zhuo Chen, Jeff Z. Pan, Zhiquan Ye, Zonggang Yuan, Yantao Jia, and Huajun Chen. Ontozsl: Ontology-enhanced zero-shot learning. In Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, editors, *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 3325–3336. ACM / IW3C2, 2021.
- [136] Zheng Liu, Zidong Jiang, and Feng Wei. OD-GCN object detection by knowledge graph with GCN. *CoRR*, abs/1908.04385, 2019.
- [137] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8303–8311. AAAI Press, 2019.
- [138] Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. From large scale image categorization to entry-level categories. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 2768–2775. IEEE Computer Society, 2013.
- [139] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1358–1367. IEEE Computer Society, 2017.
- [140] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 20–28. IEEE Computer Society, 2017.
- [141] Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. Few-shot image recognition with knowledge transfer. In *2019 IEEE/CVF International Conference on*

- Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 441–449. IEEE, 2019.
- [142] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2021–2029. IEEE Computer Society, 2017.
- [143] Xiaodan Liang, Zhiting Hu, Hao Zhang, Liang Lin, and Eric P. Xing. Symbolic graph reasoning meets convolutions. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1858–1868, 2018.
- [144] Vincent P. A. Lonij, Amrith Rawat, and Maria-Irina Nicolae. Open-world visual recognition using knowledge graphs. *CoRR*, abs/1708.08310, 2017.
- [145] Fang Yuan, Zhe Wang, Jie Lin, Luis Fernando D’Haro, Kim Jung Jae, Zeng Zeng, and Vijay Chandrasekhar. End-to-end video classification with knowledge graphs. *CoRR*, 2017.
- [146] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. Explicit knowledge-based reasoning for visual question answering. In *IJCAI*, 2017.
- [147] Michael Färber, Basil Ell, Carsten Menne, and Achim Rettinger. A comparative survey of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web Journal*, 1(1):1–5, 2015.
- [148] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 1995.
- [149] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, 2007.
- [150] Denny Vrandečić. Wikidata: a new platform for collaborative data collection. In *Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume)*, pages 1063–1064. ACM, 2012.
- [151] Filip Ilievski, Pedro A. Szekely, and Daniel Schwabe. Commonsense knowledge in wikidata. *CoRR*, abs/2008.08114, 2020.

- [152] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press, 2017.
- [153] Filip Ilievski, Pedro Szekely, and Bin Zhang. Cskg: The commonsense knowledge graph. *Extended Semantic Web Conference (ESWC)*, 2021.
- [154] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 951–958. IEEE Computer Society, 2009.
- [155] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 647–655. JMLR.org, 2014.
- [156] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [157] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9):2251–2265, 2019.
- [158] Technical report.
- [159] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.
- [160] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(7):1425–1438, 2016.
- [161] Genevieve Patterson and James Hays. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 2751–2758. IEEE Computer Society, 2012.

- [162] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(3):453–465, 2014.
- [163] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, and Li Fei-Fei. Fine-grained car detection for visual census estimation. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4502–4508. AAAI Press, 2017.
- [164] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.
- [165] Bowen Zhang, Hexiang Hu, Vihan Jain, Eugene Ie, and Fei Sha. Learning to represent image and text with denotation graph. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 823–839. Association for Computational Linguistics, 2020.
- [166] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, volume 11205 of *Lecture Notes in Computer Science*, pages 397–414. Springer, 2018.
- [167] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014.
- [168] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151, 2011.
- [169] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna

- Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2556–2565. Association for Computational Linguistics, 2018.
- [170] Edwin G. Ng, Bo Pang, Piyush Sharma, and Radu Soricut. Understanding guided image captioning performance across domains. In Arianna Bisazza and Omri Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning, CoNLL 2021, Online, November 10-11, 2021*, pages 183–193. Association for Computational Linguistics, 2021.
- [171] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017.
- [172] Ruotian Luo, Ning Zhang, Bohyung Han, and Linjie Yang. Context-aware zero-shot recognition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11709–11716. AAAI Press, 2020.
- [173] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.
- [174] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 2019.
- [175] Hao Wang, Pu Lu, Hui Zhang, Mingkun Yang, Xiang Bai, Yongchao Xu, Mengchao He, Yongpan Wang, and Wenyu Liu. All you need is boundary: Toward arbitrary-shaped text spotting. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12160–12167. AAAI Press, 2020.
- [176] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9661–9669, October 2021.

- [177] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15262–15271. Computer Vision Foundation / IEEE, 2021.
- [178] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *CoRR*, abs/2006.16241, 2020.
- [179] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *CoRR*, abs/2004.07780, 2020.
- [180] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3630–3638, 2016.
- [181] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.
- [182] Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.
- [183] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [184] Antonio Torralba, Robert Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1958–1970, 2008.
- [185] Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. TADAM: task dependent adaptive metric for improved few-shot learning. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information*

- Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 719–729, 2018.
- [186] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*, volume 6314 of *Lecture Notes in Computer Science*, pages 213–226. Springer, 2010.
- [187] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5385–5394. IEEE Computer Society, 2017.
- [188] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *CoRR*, abs/1710.06924, 2017.
- [189] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic mapping studies in software engineering. In Giuseppe Visaggio, Maria Teresa Baldassarre, Stephen G. Linkman, and Mark Turner, editors, *12th International Conference on Evaluation and Assessment in Software Engineering, EASE 2008, University of Bari, Italy, 26-27 June 2008, Workshops in Computing*. BCS, 2008.
- [190] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. Guidelines for conducting systematic mapping studies in software engineering: An update. *Inf. Softw. Technol.*, 64:1–18, 2015.
- [191] Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In Martin J. Shepperd, Tracy Hall, and Ingunn Myrtrveit, editors, *18th International Conference on Evaluation and Assessment in Software Engineering, EASE '14, London, England, United Kingdom, May 13-14, 2014*, pages 38:1–38:10. ACM, 2014.
- [192] Jia Deng, Jonathan Krause, Alexander C. Berg, and Fei-Fei Li. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3450–3457. IEEE Computer Society, 2012.
- [193] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7450–7459. Computer Vision Foundation / IEEE, 2019.

- [194] Ruslan Salakhutdinov, Antonio Torralba, and Joshua B. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 1481–1488. IEEE Computer Society, 2011.
- [195] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7239–7248. IEEE Computer Society, 2018.
- [196] Chenhan Jiang, Hang Xu, Xiaodan Liang, and Liang Lin. Hybrid knowledge routed modules for large-scale object detection. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1559–1570, 2018.
- [197] Mrigank Rochan and Yang Wang. Weakly supervised localization of novel objects using appearance transfer. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4315–4324. IEEE Computer Society, 2015.
- [198] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3010–3019. IEEE Computer Society, 2017.
- [199] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P. Xing. Rethinking knowledge graph propagation for zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 11487–11496. Computer Vision Foundation / IEEE, 2019.
- [200] Riquan Chen, Tianshui Chen, Xiaolu Hui, Hefeng Wu, Guanbin Li, and Liang Lin. Knowledge graph transfer network for few-shot recognition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 10575–10582. AAAI Press, 2020.
- [201] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1004–1013. Computer Vision Foundation / IEEE Computer Society, 2018.

- [202] Jingjing Li, Mengmeng Jing, Ke Lu, Zhengming Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7402–7411. Computer Vision Foundation / IEEE, 2019.
- [203] Maunil R. Vyas, Hemanth Venkateswara, and Sethuraman Panchanathan. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 70–86. Springer, 2020.
- [204] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5177–5186. Computer Vision Foundation / IEEE, 2019.
- [205] Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. Zero-shot learning with semantic output codes. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pages 1410–1418. Curran Associates, Inc., 2009.
- [206] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.
- [207] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4166–4174. IEEE Computer Society, 2015.
- [208] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4447–4456. IEEE Computer Society, 2017.
- [209] Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.

- [210] Mirantha Jayathilaka, Tingting Mu, and Uli Sattler. Ontology-based n-ball concept embeddings informing few-shot image classification. In Mehwish Alam, Mehdi Ali, Paul Groth, Pascal Hitzler, Jens Lehmann, Heiko Paulheim, Achim Rettinger, Harald Sack, Afshin Sadeghi, and Volker Tresp, editors, *Machine Learning with Symbolic Methods and Knowledge Graphs co-located with European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2021), Virtual, September 17, 2021*, volume 2997 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.
- [211] Sebastian Monka, Lavdim Halilaj, Stefan Schmid, and Achim Rettinger. Learning visual models using a knowledge graph as a trainer. In Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam M. Barnaghi, Armin Haller, Mauro Dragoni, and Harith Alani, editors, *The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24-28, 2021, Proceedings*, volume 12922 of *Lecture Notes in Computer Science*, pages 357–373. Springer, 2021.
- [212] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, volume 9911 of *Lecture Notes in Computer Science*, pages 67–84. Springer, 2016.
- [213] Yongxin Yang and Timothy M. Hospedales. A unified perspective on multi-domain and multi-task learning. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [214] Abhinaba Roy, Deepanway Ghosal, Erik Cambria, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. Improving zero shot learning baselines with commonsense knowledge. *CoRR*, abs/2012.06236, 2020.
- [215] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-view zero-shot learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(11):2332–2345, 2015.
- [216] Lei Jimmy Ba, Kevin Swersky, Sanja Fidler, and Ruslan Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4247–4255. IEEE Computer Society, 2015.
- [217] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5327–5336. IEEE Computer Society, 2016.

- [218] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3591–3600. IEEE Computer Society, 2017.
- [219] Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Transferable contrastive network for generalized zero-shot learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9764–9773. IEEE, 2019.
- [220] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society, 2015.
- [221] Yuxing Tang, Josiah Wang, Xiaofang Wang, Boyang Gao, Emmanuel Dellandréa, Robert J. Gaizauskas, and Liming Chen. Visual and semantic knowledge transfer for large scale semi-supervised object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):3045–3058, 2018.
- [222] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantic aligned pre-training for vision-language tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer, 2020.
- [223] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *CoRR*, abs/2006.16934, 2020.
- [224] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. *CoRR*, abs/2102.01987, 2021.
- [225] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. *CoRR*, abs/2010.00747, 2020.
- [226] B. Shahbaba and R. Neal. Improving classification when a class hierarchy is available using a hierarchy-based prior. *Bayesian Analysis*, 2:221–237, 2005.
- [227] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL, 2014.

- [228] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017.
- [229] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *Int. J. Comput. Vis.*, 127(3):302–321, 2019.
- [230] Tomáš Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119, 2013.
- [231] Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. Reasoning with neural tensor networks for knowledge base completion. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 926–934, 2013.
- [232] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1199–1208. Computer Vision Foundation / IEEE Computer Society, 2018.
- [233] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 953–962. Computer Vision Foundation / IEEE, 2021.
- [234] Lei Zhang. Transfer adaptation learning: A decade survey. *CoRR*, abs/1903.04687, 2019.
- [235] Karl R. Weiss, Taghi M. Khoshgoftaar, and Dingding Wang. A survey of transfer learning. *J. Big Data*, 3:9, 2016.
- [236] Mohsen Kaboli. A Review of Transfer Learning Algorithms. Research report, Technische Universität München, August 2017. Transfer Learning Algorithms.
- [237] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

- [238] Jing Zhang, Wanqing Li, Philip Ogunbona, and Dong Xu. Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective. *ACM Comput. Surv.*, 52(1):7:1–7:38, 2019.
- [239] Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2):13:1–13:37, 2019.
- [240] Filippos Gouidis, Alexandros Vassiliades, Theodore Patkos, Antonis A. Argyros, Nick Bassiliades, and Dimitris Plexousakis. A review on intelligent object perception methods combining knowledge-based reasoning and machine learning. In Andreas Martin, Knut Hinkelmann, Hans-Georg Fill, Aurorea Gerber, Doug Lenat, Reinhard Stolle, and Frank van Harmelen, editors, *Proceedings of the AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice, AAAI-MAKE 2020, Palo Alto, CA, USA, March 23-25, 2020, Volume I*, volume 2600 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.
- [241] Zheyuan Ding, Li Yao, Bin Liu, and Junfeng Wu. Review of the application of ontology in the field of image object recognition. In *Proceedings of the 11th International Conference on Computer Modeling and Simulation, ICCMS 2019, North Rockhampton, QLD, Australia, January 16-19, 2019*, pages 142–146. ACM, 2019.
- [242] Giuseppe Futia and Antonio Vetrò. On the integration of knowledge graphs into deep learning models for a more comprehensible AI - three challenges for future research. *Inf.*, 11(2):122, 2020.
- [243] Jiaoyan Chen, Freddy Lecue, Jeff Pan, Ian Horrocks, and Huajun Chen. Knowledge-based Transfer Learning Explanation. In *KR2018 - Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference*, Tempe, United States, October 2018.
- [244] Shruthi Chari, Daniel M. Gruen, Oshani Seneviratne, and Deborah L. McGuinness. Directions for explainable knowledge-enabled systems. In Ilaria Tiddi, Freddy Lécué, and Pascal Hitzler, editors, *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, volume 47 of *Studies on the Semantic Web*, pages 245–261. IOS Press, 2020.
- [245] Arne Seeliger, Matthias Pfaff, and Helmut Krcmar. Semantic web technologies for explainable machine learning models: A literature review. In Elena Demidova, Stefan Dietze, John G. Breslin, Simon Gottschalk, Philipp Cimiano, Basil Ell, Agnieszka Lawrynowicz, Laura Moss, and Axel-Cyrille Ngonga Ngomo, editors, *Joint Proceedings of the 6th International Workshop on Dataset PROFILing and Search & the 1st Workshop on Semantic Explainability co-located with the 18th International Semantic Web Conference (ISWC 2019)*, Auckland, New Zealand,

- October 27, 2019, volume 2465 of *CEUR Workshop Proceedings*, pages 30–45. CEUR-WS.org, 2019.
- [246] Jiaoyan Chen, Yuxia Geng, Zhuo Chen, Ian Horrocks, Jeff Z. Pan, and Huajun Chen. Knowledge-aware zero-shot learning: Survey and perspective. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4366–4373. ijcai.org, 2021.
- [247] Somak Aditya, Yezhou Yang, and Chitta Baral. Integrating knowledge and reasoning in image understanding. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6252–6259. ijcai.org, 2019.
- [248] Johannes Stalkamp, Marc Schlipf, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 2012.
- [249] Yi Yang, Hengliang Luo, Huarong Xu, and Fuchao Wu. Towards real-time traffic sign detection and classification. *IEEE Trans. Intell. Transp. Syst.*, 2016.
- [250] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
- [251] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016.
- [252] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, 2018.
- [253] Maximilian Nickel, Lorenzo Rosasco, and Tomaso A. Poggio. Holographic embeddings of knowledge graphs. In Dale Schuurmans and Michael P. Wellman, editors, *AAAI*, 2016.
- [254] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1811–1818. AAAI Press, 2018.
- [255] Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Q. Phung. A novel embedding model for knowledge base completion based on convolutional neural network. In *NAACL-HLT*, 2018.

- [256] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [257] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, and Evan Dorundo et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *CoRR*, 2020.
- [258] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.
- [259] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CoRR*, 2019.
- [260] Princeton University. *About WordNet*, 2010.
- [261] Mark van Assem, Antoine Isaac, and Jacco von Osssenbruggen. WordNet 3.0 in RDF, 2010.
- [262] W. X. Wilcke, Peter Bloem, Victor de Boer, R. H. van t Veer, and F. A. H. van Harmelen. End-to-end entity classification on multimodal knowledge graphs. *CoRR*, 2020.
- [263] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, 2018.
- [264] Yuxia Geng, Jiaoyan Chen, Ernesto Jiménez-Ruiz, and Huajun Chen. Human-centric transfer learning explanation via knowledge graph [extended abstract]. *CoRR*, 2019.
- [265] Freddy Lécué, Jiaoyan Chen, Jeff Z. Pan, and Huajun Chen. Knowledge-based explanations for transfer learning. *Studies on the Semantic Web*. 2020.
- [266] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.
- [267] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.
- [268] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In *ECCV*, 2020.
- [269] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 2007.
- [270] Tim Lauer, Philipp Schmidt, and Melissa Vö. The role of contextual materials in object recognition. *Scientific Reports*, 2021.
- [271] Antonio Torralba. Contextual priming for object detection. *Int. J. Comput. Vis.*, 2003.

- [272] Eva Rafetseder, Sarah Schuster, Stefan Hawelka, Martin Doherty, Britt Anderson, James Danckert, and Elisabeth Stöttinger. Children struggle beyond preschool-age in a continuous version of the ambiguous figures task. *Psychological Research*, page 828–841, 2021.
- [273] Deborah Chambers and Daniel Reisberg. Can mental images be ambiguous? *J. Exp. Psychol. Human Perception Perform.*, 11(3):317–328, 1985.
- [274] Joseph Jastrow. *Fact and fable in psychology*. D Appleton & Company, New York, 1900.
- [275] F Attneave. Multistability in perception. *Sci. Am.*, 225(6):63–71, December 1971.
- [276] P. Brugger and S. Brugger. The easter bunny in october: Is it disguised as a duck? *Perceptual and motor skills*, 76,2, 1993.
- [277] Michael E. R. Nicholls, Owen Churches, and Tobias Loetscher. Perception of an ambiguous figure is affected by own-age social biases. *Sci Rep* 8, 12661, 2018.
- [278] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94 2:115–147, 1987.
- [279] Alex Martin. Grapes—grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic Bulletin & Review*, 2016.
- [280] Stefania Bracci, Nicky Daniels, and Hans Op de Beeck. Task Context Overrides Object- and Category-Related Representational Content in the Human Parietal Cortex. *Cerebral Cortex*, 2017.
- [281] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *7th International Conference on Learning Representations, ICLR*, 2019.
- [282] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition CVPR*, 2020.
- [283] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Computer Vision - ECCV - 15th European Conference, Proceedings, Part XVI*, 2018.
- [284] Xi Yang, Jie Yan, Wen Wang, Shaoyi Li, Bo Hu, and Jian Lin. Brain-inspired models for visual object recognition: an overview. *Artificial Intelligence Review*, 2022.
- [285] M.F. Bonner and R.A. Epstein. Object representations in the human brain reflect the co-occurrence statistics of vision and language. In *Nat Commun* 12, 2021.

- [286] Thomas N. Kipf and Max Welling. Variational graph auto-encoders. *CoRR*, abs/1611.07308, 2016.
- [287] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.
- [288] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.
- [289] Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Contextual models for object detection using boosted random fields. In *Neural Inf. Processing Systems NIPS*, 2004.
- [290] Kevin Wu, Eric Wu, and Gabriel Kreiman. Learning scene gist with convolutional neural networks to improve object recognition. In *52nd Annual Conference on Information Sciences and Systems CISS*, 2018.
- [291] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In *Advances in Neural Information Processing Systems: Annual Conf. on Neural Information Processing Systems*, 2018.
- [292] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.
- [293] Qihang Yu, Yingda Xia, Yutong Bai, Yongyi Lu, Alan Yuille, and Wei Shen. Glance-and-Gaze Vision Transformer. (NeurIPS), 2021.
- [294] Peng Gao, Jiasen Lu, Hongsheng Li, Roozbeh Mottaghi, and Aniruddha Kembhavi. Container: Context aggregation networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems*, 2021.
- [295] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Geometric context from a single image. In *International Conference on Computer Vision ICCV*. Computer Society, 2005.
- [296] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Computer Vision - ECCV - 12th European Conference on Computer Vision*, 2012.
- [297] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016.

- [298] Mengmi Zhang, Jiashi Feng, Karla Montejo, Joseph Kwon, Joo Hwee Lim, and Gabriel Kreiman. Lift-the-flap: Context reasoning using object-centered graphs. *CoRR*, abs/1902.00163, 2019.
- [299] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2018.
- [300] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017.
- [301] Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. Learning structured inference neural networks with label relations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Society, 2016.
- [302] Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, 2016.
- [303] Yongzhi Li, Duo Zhang, and Yadong Mu. Visual-semantic matching by exploring high-order attention and distraction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2020.
- [304] Eloi Zablocki, Patrick Bordes, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. Context-aware zero-shot learning for object recognition. In *Proceedings of the 36th International Conference on Machine Learning ICML*, 2019.
- [305] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Vision Foundation, 2020.
- [306] Simon Werner, Achim Rettinger, Lavdim Halilaj, and Jürgen Lüttin. RETRA: recurrent transformers for learning temporally contextualized knowledge graph embeddings. In *The Semantic Web - 18th International Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings*, 2021.
- [307] Zhiyuan Ning, Ziyue Qiao, Hao Dong, Yi Du, and Yuanchun Zhou. Lightcake: A lightweight framework for context-aware knowledge graph embedding. In *Advances in Knowledge Discovery and Data Mining - 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11-14, 2021, Proceedings, Part III*, 2021.

- [308] Haoyu Wang, Vivek Kulkarni, and William Yang Wang. Dolores: Deep contextualized knowledge graph embeddings. In *Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020*, 2020.
- [309] Quan Wang, Pingping Huang, Haifeng Wang, Songtai Dai, Wenbin Jiang, Jing Liu, Yajuan Lyu, Yong Zhu, and Hua Wu. Coke: Contextualized knowledge graph embedding. 2019.
- [310] Suzana Herculano-Houzel. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences*, 109(supplement_1):10661–10668, June 2012.
- [311] Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. RML: A generic language for integrated RDF mappings of heterogeneous data. In Christian Bizer, Tom Heath, Sören Auer, and Tim Berners-Lee, editors, *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014*, volume 1184 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.
- [312] Natthawut Kertkeidkachorn and Ryutaro Ichise. An automatic knowledge graph creation framework from natural language text. *IEICE Trans. Inf. Syst.*, 101-D(1):90–98, 2018.
- [313] Danilo Dessì, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, and Enrico Motta. Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain. *Future Gener. Comput. Syst.*, 116:253–264, 2021.
- [314] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.
- [315] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*. MIT Press, 2001.
- [316] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: online learning of social representations. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*. ACM, 2014.
- [317] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. ACM, 2016.

- [318] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. ACM, 2016.
- [319] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002.
- [320] Petar Velickovic, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. Deep graph infomax. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.
- [321] Chen Xing, Negar Rostamzadeh, Boris N. Oreshkin, and Pedro O. Pinheiro. Adaptive cross-modal few-shot learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4848–4858, 2019.
- [322] Olaf Hartig. Foundations of rdf \star and sparql \star (an alternative approach to statement-level metadata in RDF). In *Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web, Montevideo, Uruguay, June 7-9, 2017*, 2017.
- [323] Olaf Hartig. Foundations to query labeled property graphs using SPARQL. In *Joint Proceedings of the 1st International Workshop On Semantics For Transport and the 1st International Workshop on Approaches for Making Data Interoperable co-located with 15th Semantics Conference (SEMANTiCS 2019), Karlsruhe, Germany, September 9, 2019*, 2019.
- [324] Jonathan Hayes and Claudio Gutiérrez. Bipartite graphs as intermediate model for RDF. In *The Semantic Web - ISWC 2004: Third International Semantic Web Conference, Hiroshima, Japan, November 7-11, 2004. Proceedings*, pages 47–61, 2004.
- [325] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In Vera Kurková, Yannis Manolopoulos, Barbara Hammer, Lazaros S. Iliadis, and Ilias Maglogiannis, editors, *Artificial Neural Networks and Machine Learning - ICANN 2018 - 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III*, volume 11141 of *Lecture Notes in Computer Science*, pages 270–279. Springer, 2018.
- [326] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 554–561. IEEE Computer Society, 2013.

- [327] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.*, 41(6):391–407, 1990.
- [328] Mingsheng Long, Jianmin Wang, Guiguang Ding, Dou Shen, and Qiang Yang. Transfer learning with graph co-regularization. *IEEE Trans. Knowl. Data Eng.*, 26(7):1805–1818, 2014.
- [329] Zhengming Ding, Sheng Li, Ming Shao, and Yun Fu. Graph adaptive knowledge transfer for unsupervised domain adaptation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, volume 11206 of *Lecture Notes in Computer Science*, pages 36–52. Springer, 2018.
- [330] Hanrui Wu, Yuguang Yan, Yuzhong Ye, Michael K. Ng, and Qingyao Wu. Geometric knowledge embedding for unsupervised domain adaptation. *Knowl. Based Syst.*, 191:105155, 2020.
- [331] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 6034–6042. IEEE Computer Society, 2016.
- [332] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5716–5726. IEEE Computer Society, 2017.
- [333] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G. Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4893–4902. Computer Vision Foundation / IEEE, 2019.
- [334] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV*, volume 12359 of *Lecture Notes in Computer Science*, pages 266–282. Springer, 2020.
- [335] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society, 2015.

- [336] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.
- [337] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. IEEE, 2020.
- [338] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*, pages 873–882. The Association for Computer Linguistics, 2012.
- [339] Zhengming Ding, Ming Shao, and Yun Fu. Latent low-rank transfer subspace learning for missing modality recognition. In Carla E. Brodley and Peter Stone, editors, *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, pages 1192–1198. AAAI Press, 2014.
- [340] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 343–351, 2016.
- [341] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society, 2015.
- [342] Maximilian Nickel, Lorenzo Rosasco, and Tomaso A. Poggio. Holographic embeddings of knowledge graphs. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 1955–1961. AAAI Press, 2016.
- [343] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 809–816. Omnipress, 2011.

- [344] Deepak Nathani, Jatin Chauhan, Charu Sharma, and Manohar Kaul. Learning attention-based embeddings for relation prediction in knowledge graphs. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4710–4723. Association for Computational Linguistics, 2019.
- [345] Sebastian Monka, Lavdim Halilaj, Stefan Schmid, and Achim Rettinger. *Contrakg: Contrastive-based transfer learning for visual object recognition using knowledge graphs*. 2020.
- [346] H. Hanke and D. Knees. A phase-field damage model based on evolving microstructure. *Asymptotic Analysis*, 101:149–180, 2017.
- [347] E. Lefever. A hybrid approach to domain-independent taxonomy learning. *Applied Ontology*, 11(3):255–278, 2016.
- [348] P.S. Meltzer, A. Kallioniemi, and J.M. Trent. Chromosome alterations in human solid tumors. In B. Vogelstein and K.W. Kinzler, editors, *The Genetic Basis of Human Cancer*, pages 93–113. McGraw-Hill, New York, 2002.
- [349] P.R. Murray, K.S. Rosenthal, G.S. Kobayashi, and M.A. Pfaller. *Medical Microbiology*. Mosby, St. Louis, 4th edition, 2002.
- [350] E. Wilson. *Active vibration analysis of thin-walled beams*. PhD thesis, University of Virginia, 1991.
- [351] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.
- [352] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015.
- [353] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.
- [354] Tom Michael Mitchell, S. V. Shinkareva, A. Carlson, Kai min Chang, Vicente L. Malave, R. Mason, and M. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 2008.
- [355] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.
- [356] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

- [357] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.
- [358] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017.
- [359] Lei Jimmy Ba, Kevin Swersky, Sanja Fidler, and Ruslan Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*, 2015.
- [360] Maryam Vaziri-Pashkam and Yaoda Xu. An Information-Driven 2-Pathway Characterization of Occipitotemporal and Posterior Parietal Visual Object Representations. *Cerebral Cortex*, 29(5):2034–2050, 04 2018.
- [361] Aude Oliva, Jeremy M. Wolfe, and Helga C. Arsenio. Panoramic search: The interaction of memory and vision in search through a familiar scene. *Journal of Experimental Psychology: Human Perception and Performance*, 2004.
- [362] Melissa Le-Hoa Võ and Jeremy M. Wolfe. The role of memory for visual search in scenes. *Annals of the New York Academy of Sciences*, 1339(1):72–81, 2015.
- [363] Michelle R. Greene, Christopher A. Baldassano, Andre Esteva, Diane M. Beck, and Li Fei-Fei. Visual scenes are categorized by function. *Journal of experimental psychology. General*, 145 1:82–94, 2016.
- [364] Kyle E. Mathewson. Duck eats rabbit: Exactly which type of relational phrase can disambiguate the perception of identical side by side ambiguous figures? *Perception*, 47(4):466–469, 2018. PMID: 29402155.
- [365] Sebastian Monka, Lavdim Halilaj, and Achim Rettinger. A survey on visual transfer learning using knowledge graphs. *Semantic Web*, 13(3):477–510, 2022.
- [366] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [367] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.
- [368] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*. 2019.

- [369] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, 2019.
- [370] Sebastian Monka, Lavdim Halilaj, and Achim Rettinger. Context-driven visual object recognition based on knowledge graphs. In Ulrike Sattler, Aidan Hogan, C. Maria Keet, Valentina Presutti, João Paulo A. Almeida, Hideaki Takeda, Pierre Monnin, Giuseppe Pirrò, and Claudia d’Amato, editors, *The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings*, volume 13489 of *Lecture Notes in Computer Science*, pages 142–160. Springer, 2022.
- [371] Michael L. Littman, Ifeoma Ajunwa, Guy Berger, Craig Boutilier, Morgan Currie, Finale Doshi-Velez, Gillian K. Hadfield, Michael C. Horowitz, Charles Isbell, Hiroaki Kitano, Karen Levy, Terah Lyons, Melanie Mitchell, Julie Shah, Steven Sloman, Shannon Vallor, and Toby Walsh. Gathering strength, gathering storms: The one hundred year study on artificial intelligence (AI100) 2021 study panel report. *CoRR*, abs/2210.15767, 2022.
- [372] George Boole. *The Laws of Thought*. London: MacMillan and Co, 1988.
- [373] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multi-task learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [374] Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M Jacobs, Jens Bölte, and Andrea Böhl. The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in german. *Exp. Psychol.*, 58(5):412–424, 2011.
- [375] Victoria Fromkin, Robert Rodman, and Nina Hyams. *An Introduction to Language*. Wadsworth, Cengage Learning, South Melbourne, Victoria, 9. ed., international student ed edition, 2011.
- [376] Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? A controlled study for representation learning. *CoRR*, abs/2207.07635, 2022.