# Few-Shot and Zero-Shot Learning for Information Extraction

Jiaying Gong

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Hoda M. Eldardiry, Chair
Jia-Bin Huang
Lifu Huang
Ismini Lourentzou
Dawei Zhou

April 25, 2024
Blacksburg, Virginia

Keywords: Information Extraction, Few-Shot Learning, Zero-Shot Learning.

# Few-Shot and Zero-Shot Learning for Information Extraction

Jiaying Gong

(ABSTRACT)

Information extraction aims to automatically extract structured information from unstructured texts. Supervised information extraction requires large quantities of labeled training data, which is time-consuming and labor-intensive. This dissertation focuses on information extraction, especially relation extraction and attribute-value extraction in e-commerce, with few labeled (few-shot learning) or even no labeled (zero-shot learning) training data. We explore multi-source auxiliary information and novel learning techniques to integrate semantic auxiliary information with the input text to improve few-shot learning and zero-shot learning.

For zero-shot and few-shot relation extraction, the first method explores the existing data statistics and leverages auxiliary information including labels, synonyms of labels, keywords, and hypernyms of name entities to enable zero-shot learning for the unlabeled data. We build an automatic hypernym extraction framework to help acquire hypernyms of different entities directly from the web. The second method explores the relations between seen classes and new classes. We propose a prompt-based model with semantic knowledge augmentation to recognize new relation triplets under the zero-shot setting. In this method, we transform the problem of zero-shot learning into supervised learning with the generated augmented data for new relations. We design the prompts for training using the auxiliary information based on an external knowledge graph to integrate semantic knowledge learned from seen relations. The third work utilizes auxiliary information from images to enhance few-shot learning. We propose a multi-modal few-shot relation extraction model that leverages both textual and visual semantic information to learn a multi-modal representation jointly. To supplement the missing contexts in text, this work integrates both local features (object-level) and global features (pixel-level) from different modalities through image-guided attention, object-guided attention, and hybrid feature attention to solve the problem of sparsity and noise.

We then explore the few-shot and zero-shot aspect (attribute-value) extraction in the e-commerce application field. The first work studies the multi-label few-shot learning by leveraging the auxiliary information of anchor (label) and category description based on the prototypical networks, where the hybrid attention helps alleviate ambiguity and capture more informative semantics by calculating both the label-relevant and query-related weights. A dynamic threshold is learned by integrating the semantic information from support and query sets to achieve multi-label inference. The second work explores multi-label zero-shot learning via semi-inductive link prediction of the heterogeneous hypergraph. The heterogeneous hypergraph is built with higher-order relations (generated by the auxiliary information of user behavior data and product inventory data) to capture the complex and interconnected relations between users and the products.

# Few-Shot and Zero-Shot Learning for Information Extraction

Jiaying Gong

(GENERAL AUDIENCE ABSTRACT)

Information extraction is the process of automatically extracting structured information from unstructured sources, such as plain text documents, web pages, images, and so on. In this dissertation, we will first focus on general relation extraction, which aims at identifying and classifying semantic relations between entities. For example, given the sentence 'Peter was born in Manchester.' in the newspaper, structured information (Peter, place of birth, Manchester) can be extracted. Then, we focus on attribute-value (aspect) extraction in the application field, which aims at extracting attribute-value pairs from product descriptions or images on e-commerce websites. For example, given a product description or image of a handbag, the brand (i.e. brand: Chanel), color (i.e. color: black), and other structured information can be extracted from the product, which provides a better search and recommendation experience for customers.

With the advancement of deep learning techniques, machines (models) trained with large quantities of example input data and the corresponding desired output data, can perform automatic information extraction tasks with high accuracy. Such example input data and the corresponding desired output data are also named annotated data. However, across technological innovation and social change, new data (i.e. articles, products, etc.) is being generated continuously. It is difficult, time-consuming, and costly to annotate large quantities of new data for training. In this dissertation, we explore several different methods to help the model achieve good performance with only a few (few-shot learning) or even no labeled data (zero-shot learning) for training.

Humans are born with no prior knowledge, but they can still recognize new information based on their existing knowledge by continuously learning. Inspired by how human beings learn new knowledge, we explore different auxiliary information that can benefit few-shot and zero-shot information extraction. We studied the auxiliary information from existing data statistics, knowledge graphs, corresponding images, labels, user behavior data, product inventory data, optical characters, etc. We enable few-shot and zero-shot learning by adding auxiliary information to the training data. For example, we study the data statistics of both labeled and unlabeled data. We use data augmentation and prompts to generate training samples for no labeled data. We utilize graphs to learn general patterns and representations that can potentially transfer to unseen nodes and relations. This dissertation provides the exploration of how utilizing the above different auxiliary information to help improve the performance of information extraction with few annotated or even no annotated training data.

# Acknowledgments

First and foremost, I would like to express my sincerest gratitude to my advisor, Dr. Eldardiry, who has supported me a lot throughout my Ph.D. journey with guidance, encouragement, patience, and support. She is a very kind and supportive advisor, not only in research but also in my career and life. It is her professional guidance, unwavering patience, and continuous encouragement that helped me stay confident and positive in every challenge I have met during my Ph.D. years. She has made lots of efforts to guide me toward new research directions, brainstorming, and improving my skills in presentations, teaching, academic writing, and independent research.

I would also like to thank all my committee members: Dr. Jia-Bin Huang, Dr. Lifu Huang, Dr. Ismini Lourentzou, and Dr. Dawei Zhou, who have provided valuable suggestions and important guidance for the completion of my dissertation. Special thanks to Dr. Jia-Bin Huang for the motivation of my research and inspiration for dissertation topics.

Besides, I would like to express my appreciation to my colleagues in the Machine Learning Laboratory: Chenhan Yuan, Hongjie Chen, Vasanth Reddy Baddam, Ming Cheng, Afrina Tabassum, Jiazhen Hu, etc, for research discussions. I would especially like to express my thanks to my five-year roommate Jingyi Zhang, who helped me a lot during my life from the first day I came to the U.S. I cherish all the moments we spent together on research discussions, course-related projects, cooking, baking, and traveling. Thank you for bringing so many good memories and times of happiness to my life.

Finally, I would like to dedicate a special thanks to my parents for all of their love and support. They have provided me with enormous support, patience, understanding, and love throughout my Ph.D. journey.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Information extraction aims to automatically extract structured information from unstructured texts. The main goal of information extraction is to find meaningful information from the unstructured text input. The information can be name entities, relations between the name entities, events, aspect categories, attribute-value pairs, and any other pre-defined structured information. For example, relation triplet <London, capital of, England> is a structured triplet of the sentence: Patten emigrated to England, working in London. Aspect (attribute-value pair) <Brand: Chanel> can be extracted from the product description or image in e-commerce. This information is very important to the success of various downstream tasks, such as the construction of knowledge graphs, taxonomy enrichment, and recommendation systems in e-commerce, finance, healthcare, etc.

## 1.1 Motivation

Early works for information extraction use a domain-specific dictionary and rule-based methods to extract information [5, 6, 49, 140, 169]. With the development of neural networks, convolutional neural networks [205, 244], recurrent neural networks [252, 263], and graph neural networks [255, 264] are used to learn the unstructured input text patterns for structured information extraction. Even though these works have achieved promising results by taking advantage of supervised learning, they exhibit a fundamental limitation in that they all need large quantities of labeled training data, which requires lots of human effort. To save the effort for human labeling, distant supervision [143, 245] or crowdsourcing [1, 120] are used to collect more examples with labels. However, these methods are limited by the quantity (for supervised) and quality (for distant-supervised) of the training data because manually labeling the data is time-consuming and labor-intensive, and the data labeled by distant-supervision is noisy. To address the problem of insufficient high-quality labeled training data, we focus on few-shot and zero-shot learning for information extraction.

Few-shot learning or zero-shot learning requires only a few or even no labeled data for training. Different from traditional supervised learning, which requires the model to recognize the data in the training set and then generalize to the unseen test data, the goal of few-shot and zero-shot learning is to train the model learn to learn. To be more specific, few-shot learning and zero-shot learning enable a pre-trained model to generalize over new categories of data using

only a few labeled or even no labeled samples per class [211]. Few-shot and zero-shot learning have been widely used in computer vision fields such as image classification [93, 147, 197], image retrieval [8, 14, 184], object detection [9, 45, 224], gesture recognition [138, 165, 168], image captioning [25, 39, 47], and visual question answering [26, 39, 67]. Motivated by these works, we consider few-shot and zero-shot learning for information extraction in the field of natural language processing.

## 1.2 Research Questions and Hypothesis

The primary research question for this dissertation is: *How can multi-source learning improve few-shot and zero-shot information extraction?* To tackle the above challenges, this question can be further decomposed into the following research sub-questions.

- *RQ 1: How can weighted side information including keywords, hypernyms of name entities, and labels help for zero-shot relation extraction?*
- *RQ 2: How can data augmentation and prompts built on an external knowledge graph integrate semantic information for zero-shot relation triplet extraction?*
- *RQ 3: How can visual information as an external source supplement the missing contexts in textual sentences for few-shot relation extraction?*
- *RQ 4: How can label description enhance the performance of multi-label few-shot attribute value extraction in e-commerce?*
- *RQ5: How can heterogeneous hypergraph with higher-order relations constructed by user behavior and product inventory data improve zero-shot attribute value prediction?*

Accordingly, the central hypothesis of this research is that *the proposed different multi-source auxiliary information will improve the performance of information extraction with few labeled or even no labeled training data and the extracted information will benefit downstream tasks and practical applications.*

## 1.3 Research Issues

Although there exist many works on few-shot and zero-shot learning in the computer vision field, there are still many challenges in information extraction in the natural language processing field: (1) In the computer vision field, auxiliary information from text is always used as prior knowledge for zero-shot image-related tasks such as image classification, image captioning, etc [25, 176]. However, in the natural language processing field, the auxiliary

information is difficult to achieve. The first major challenge is what kind of auxiliary information is effective for few-shot and zero-shot information extraction. (2) Existing approaches on few-shot and zero-shot information extraction depend on question answering models [101], text entailment [154] and label description [16]. However, these models have a strong assumption that excellent question-answering models are learned and label descriptions are accurate. The second challenge is what's the connection or relationships between seen labels and novel labels. (3) The third challenge is related to multi-modal few-shot learning. Existing few-shot information extraction models (prototypical networks [51] and siamese neural networks [243]) only use plain text or external data sources (entities [236], labels [130] and graphs [161]) for training the model. However, these methods mainly explore single-modality text-based data and may suffer a significant performance decline when texts lack contexts. Besides, fusing information without considering semantic information from different modalities (such as concatenation, and circulant fusion [220]) will have a negative impact because fusing irrelevant information may bring noise. (4) In the real world, each input instance may include several labels. The fourth challenge is how to fuse auxiliary information in a multi-label few-shot information extraction task given limited auxiliary information. For a multi-label setting, it is also challenging to predict the number of labels for each data instance. (5) The fifth challenge is related to multi-label zero-shot learning. Existing works focus on open mining models [228, 254] or generative large language models [109] to directly extract information that has been mentioned in the text or autoregressively decoded from the inputs. However, these methods need high-quality seed attribute sets and heavy computing resources. Besides, they lose the higher-order relationships of the complex and interconnected semantic information.

In this dissertation, we focus on dealing with the above-mentioned challenges in few-shot and zero-shot information extraction. The detailed statement of research issues is presented in the following subsections.

### 1.3.1 Zero-Shot Relation Classification from Side Information

To extract information with few or no labeled training data, we first think of any auxiliary information we could use to help provide more prior knowledge. We consider the semantic information from the label and hypernyms of name entities. Therefore, we propose a zero-shot learning model for relation classification (ZSLRC), which focuses on recognizing new relations with no corresponding labeled data available for training. We construct weighted side information from labels and their synonyms, hypernyms of two-name entities, and keywords from training sentences. We also build an automatic hypernym extraction framework to help us acquire hypernyms of different entities directly from the web. Finally, We modify the vanilla prototypical networks [191] utilizing the above side (auxiliary) information to extract both relations with training instances and relations without any training instances. To detect novel relations, ZSLRC decides whether the query input is in a class with training data or one without any training data based on a threshold.

### 1.3.2 Prompt-based Zero-shot Relation Classification with Semantic Knowledge Augmentation

To address the challenge of loss of the connection or relationships between seen labels and novel labels, we propose a prompt-based model with semantic knowledge augmentation (ZS-SKA) to perform zero-shot learning for relation classification. We first implement data augmentation by a word-level sentence translation to generate augmented instances with unseen relations from training instances with seen relations. Then, we design the prompts based on a knowledge graph to integrate semantic knowledge to generally infer the features of unseen relations using patterns learned from seen relations. Instead of using the real label word directly in the prompt template, we automatically search a set of appropriate label words based on the knowledge graph for each label. We calculate the distance between each appropriate label with the true label itself to help denoise the set of appropriate label words. Then, we construct virtual label words in the prompt by weighted averaging all appropriate label word candidates.

### 1.3.3 Multi-Modal Few-Shot Relation Extraction with Hybrid Visual Evidence

To address the challenge that signal modality text-based data may lose semantic information when there are no clear contexts between the two name entities described in the text, we propose a Multi-modal Few-Shot model based on Hybrid Visual Evidence (MFS-HVE) for relation extraction. The first contribution is that we crawl the images automatically for each instance in a public single modality dataset [72] to provide visual information, which can facilitate future research on multi-modal few-shot relation extraction. We generate the representations through both textual feature extractors and visual feature extractors. We consider the visual representations from both the local perspective (object level) and the global perspective (pixel level). For the methods of fusing information from different modalities, we propose a multi-modal fusion unit including image-guided attention, object-guided attention, and hybrid feature attention to integrate semantic information from different modalities at both global and local levels. Finally, we concatenate text features, image-guided features, and object-guided features through a cross-modality encoder to generate the final multi-modal representations.

### 1.3.4 Knowledge-Enhanced Multi-Label Few-Shot Product Attribute-Value Extraction

To formulate the attribute-value extraction task in a multi-label few-shot setting, we follow the minimum-including algorithm for multi-label few-shot data sampling. We propose a

Knowledge-Enhanced Attentive Framework (KEAF) based on prototypical networks, leveraging the generated label description and category information as the auxiliary information to learn more discriminative prototypes. We develop the hybrid attention mechanism, which helps reduce the noise and capture more informative semantics from the support set by calculating both the label-relevant and query-related weights. To achieve multi-label inference, a dynamic threshold is learned during the training stage by integrating the semantic information from both support and query sets.

### 1.3.5 Multi-Label Zew-Shot Product Attribute-Value Extraction

To address the challenge of complex and interconnected higher-order relations among different products and users, we propose HyperPAVE, a multi-label zero-shot model based on the heterogeneous hypergraph, to predict zero-shot attribute values. The heterogeneous hypergraph is constructed by auxiliary information from interconnected user behavior information (i.e. 'also buy', 'also view') and structured product inventory information (i.e. 'product has aspects', 'category includes products'), to include more information in the final node representations and remove user nodes in the original graph. HyperPAVE leverages a semi-inductive link prediction mechanism, which is combined with a fine-tuned BERT encoder, to obtain unseen contextual node features. HyperPAVE is updated with zero-shot products and aspects, where message-passing is conducted directly on the updated graph, ensuring the inductive inference ability.

## 1.4 Dissertation Organization

The remainder of this dissertation is organized as follows. In Chapter 2, we propose a zero-shot learning framework based on prototypical networks with side information for relation classification. The side information is built from keywords, hypernyms of name entities, and labels. In Chapter 3, we propose a prompt-based model with semantic knowledge augmentation for zero-shot relation extraction. Prompts are designed by weighted virtual label construction based on an external knowledge graph. In Chapter 4, we propose a multi-modal few-shot relation extraction model leveraging both textual and visual semantic information through visual-guided attention, object-guided attention, and hybrid feature attention, to learn a multi-modal representation jointly. In Chapter 5, we propose a knowledge-enhanced attentive framework for multi-label few-shot attribute-value extraction, leveraging the generated label description and category information. A dynamic threshold is learned by integrating semantic information to predict label counts. In Chapter 6, we propose HyperPAVE, a multi-label zero-shot attribute value extraction model leveraging inductive inference in heterogeneous hypergraphs, which captures complex and interconnected higher-order relations. Chapter 7 will summarize the main contributions of the dissertation and speculate about future research directions in few/zero-shot learning for information extraction.

# Chapter 2

# Zero-Shot Relation Classification from Side Information

This chapter introduces a zero-shot learning relation classification (ZSLRC) framework that improves on state-of-the-art by its ability to recognize novel relations that were not present in training data. The zero-shot learning approach mimics the way humans learn and recognize new concepts with no prior knowledge. To achieve this, ZSLRC uses advanced prototypical networks that are modified to utilize weighted side (auxiliary) information. ZSLRC's side information is built from keywords, hypernyms of name entities, and labels and their synonyms. ZSLRC also includes an automatic hypernym extraction framework that acquires hypernyms of various name entities directly from the web. ZSLRC improves on state-of-the-art few-shot learning relation classification methods that rely on labeled training data and is therefore applicable more widely even in real-world scenarios where some relations have no corresponding labeled examples for training. We present results using extensive experiments on two public datasets (NYT and FewRel) and show that ZSLRC significantly outperforms state-of-the-art methods on supervised learning, few-shot learning, and zero-shot learning tasks. Our experimental results also demonstrate the effectiveness and robustness of our proposed model.

## 2.1 Introduction

Relation classification aims to infer the relation between two name entities in a sentence. Supervised learning methods for relation classification have been widely used to classify relations based on training labeled data. Distant supervision or crowdsourcing has been used to collect more examples with labels and train the model for relation classification. However, these methods are limited by the quantity (for supervised) and quality (for distantly supervised) of the training data because manually labeling the data is time-consuming and labor-intensive, and data labeled by distant supervision is noisy. To overcome the problem of insufficient high-quality data, few-shot learning has been designed to require only a few labeled sentences for training. A lot of research has been done on few-shot learning for computer vision [103, 115, 239], and some work also includes few-shot learning methods for relation classification [51, 72, 87]. However, these works still require a few instances for training, and they still do not work when no training instances are available.

Some work on open information extraction (OpenIE) discovers new relationships in open-domain corpora without labeling the data [4]. OpenIE aims to extract relation phrases directly from the text. However, this technique can not effectively select meaningful relation patterns and discard irrelevant information. Besides, this technique can not discover relations if the relation's name does not appear in the given sentence. For example, OpenIE can not identify the relation of the sentence in Figure 2.1.

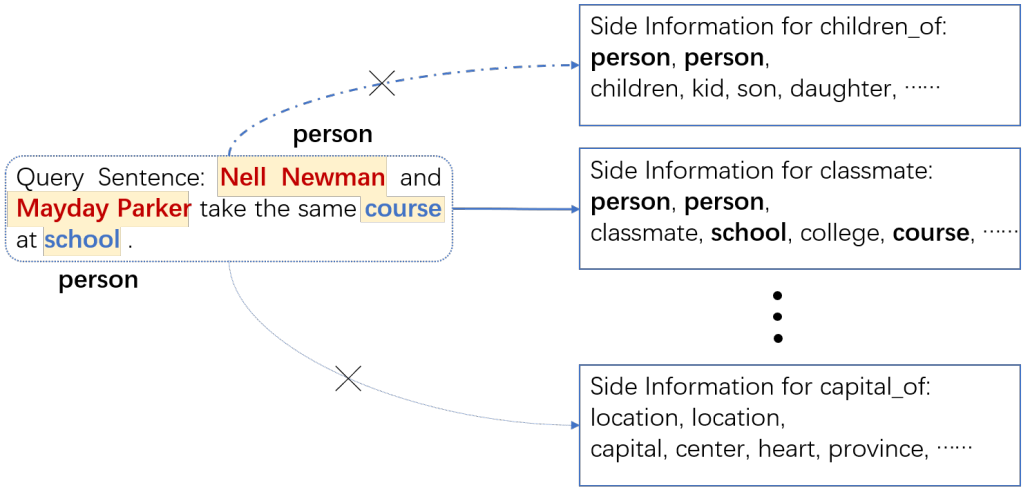


Figure 2.1: Example of relation classification based on side information.

To address the limitations mentioned above, we focus on relation classification in the context of zero-shot learning. Zero-shot learning (ZSL) is similar to the way humans learn and recognize new concepts. It is a novel learning technique that does not use any exemplars of the unseen categories during training. We propose a zero-shot learning model for relation classification (ZSLRC), which focuses on recognizing new relations with no corresponding labeled data available for training. ZSLRC is modified on prototypical networks utilizing side (auxiliary) information. We construct weighted side information from labels and their synonyms, hypernyms of two-name entities, and keywords from training sentences. The ZSL-based model can recognize new relations based on the side information available for it instead of using a collection of labeled sentences. We incorporate side information to enable our model to identify relations that never appear in the training datasets. We also build an automatic hypernym extraction framework to help us acquire hypernyms of different entities directly from the web. Details of side information construction are described in Section 2.3.2.

Figure 2.1 shows an example of how side information can be used for classifying relations. Different side information is given for different relations. The query sentence in the example has a relation of *classmate_of*, but the word classmate never appears in the sentence. We first get the two name entities *Nell Newman* and *Mayday Parker* of the sentence and extract the hypernyms of the name entities *person* and *person* based on our proposed hypernym extraction module in Section 2.3.2. In this example, relation *capital_of* is eliminated because

the hypernyms of *capital_of* should be *location* and *location*. Then we extract the keywords *course* and *school* from the query sentence and compare the distance with the keywords in the side information box. In this way, relation *children_of* is eliminated.

To make relation classification effective in real-world scenarios, we design our model with the ability to classify both relations with training instances and relations without any training instances. We modify the vanilla prototypical networks to deal with both scenarios and compare the distance between the query sentence and the weighted prototype. If the exponential of the minus distance is above a threshold, we consider the query sentence to have a new relation. For new relations identification, we take the side information embedding from the query sentence and compare the distance of it with the side information embedding of new relations. We conduct different experiments on both a noisy and a clean dataset and add different percentages of new relations to evaluate the effectiveness and robustness of our proposed model. Besides, we also evaluate our proposed model in supervised learning, few-shot learning, and zero-shot learning tasks. The results show that our proposed model outperforms other existing models in all three tasks. The contributions of this chapter can be summarized as follows:

- We propose the first approach (ZSLRC) to enable zero-shot learning on relation classification without relying on other complex models that need to be learned and assumed to be 100% accurate.
- ZSLRC uses side information including labels, keywords, and hypernyms of name entities, and it has been shown that our model can perform competitively using the weighted side information. We build an automatic hypernym extraction framework to extract hypernyms of words from the web.
- We modify prototypical networks to recognize new relations in addition to recognized previously known relations. Results show the effectiveness and robustness of our modified prototypical networks in different learning tasks.
- We demonstrate that our proposed model significantly outperforms state-of-the-art methods on supervised learning, few-shot learning, and zero-shot learning tasks. We ran extensive experiments on two datasets.

## 2.2 Related Work

**Supervised Relation Classification.** Relation Classification aims to classify relations between entities. Many existing relation classification methods are based on supervised learning, where neural networks are used to extract semantic features from text automatically. For example, convolutional neural networks (CNNs) are used to learn textual patterns [40, 121, 150, 205, 244, 265]. Recurrent neural networks (RNNs) are used to better

capture the sequential information present in the input data [149, 246, 263]. Graph neural networks (GNNs) are used to find dependencies and capture long-range relations between words [255, 264]. Although these traditional Relation Classification methods have achieved promising results by taking advantage of supervised or distantly supervised data, they exhibit a fundamental limitation since they all need large quantities of labeled training data.

**Open Relation Extraction.** Many existing approaches focus on discovering new relationships in open-domain corpora. This is because traditional supervised RC can not find new relation types due to their limited ability to classify predefined relation types. Open RE or Open information extraction (OpenIE) aims to extract relation phrases directly from the text. For example, tagging-based methods [31, 90] and clustering-based methods [139, 222] are used to discover new relation types. Other work proposed Relational Siamese Networks to transfer relational knowledge from supervised OpenRE data to calculate the similarity of unlabeled sentences for open relation clustering [222]. However, OpenRE can not effectively select meaningful relation patterns and discard irrelevant information. In the real world, methods that rely on predefined relation types are always known to lack of training data.

**Zero-shot Learning.** Zero-shot learning has been widely applied in computer vision [12, 88, 94, 110, 166, 225, 242]. Similar to zero-shot learning, few-shot learning is well-studied in the field of relation classification [37, 51, 52, 72, 240, 243]. However, compared with zero-shot learning for computer vision and few-shot learning explored in relation classification, there exists little work toward zero-shot learning in the domain of natural language processing. Some current work uses a transferable architecture to jointly represent and map event types in order to detect unseen event types [83]. Other work proposed a zero-shot learning method for relation extraction from webpages with unseen templates [133]. However, this method solves a different problem, only predicting relation types in unseen structures of webpages instead of new relation types. The most related work to zero-shot learning for relation classification uses zero-shot learning to extract unseen relation types by listing questions that define the relation's slot values [101]. However, this method requires external help, such as a question-answering dataset annotated by a human. In addition, this method assumes that (1) a good reading comprehension model is learned and that (2) all values extracted from this model are correct. In contrast, our proposed model can identify new relation types without training sentences and does not need to rely on other models. We construct weighted side information to train the model without labeled training sentences. For example, some previous works use side information from knowledge graphs or labels to lower the noise and improve performance in distantly-supervised relation classification [78, 200].

## 2.3 Methodology

In this section, we introduce the overview of the ZSLRC model. Figure 2.2 shows the architecture of zero-shot learning for relation classification. It consists of three parts: Sentence Encoder, Side Information Extraction, and Prototypical Network with Weighted Side Infor-

mation Embedding. We describe these parts in detail below.

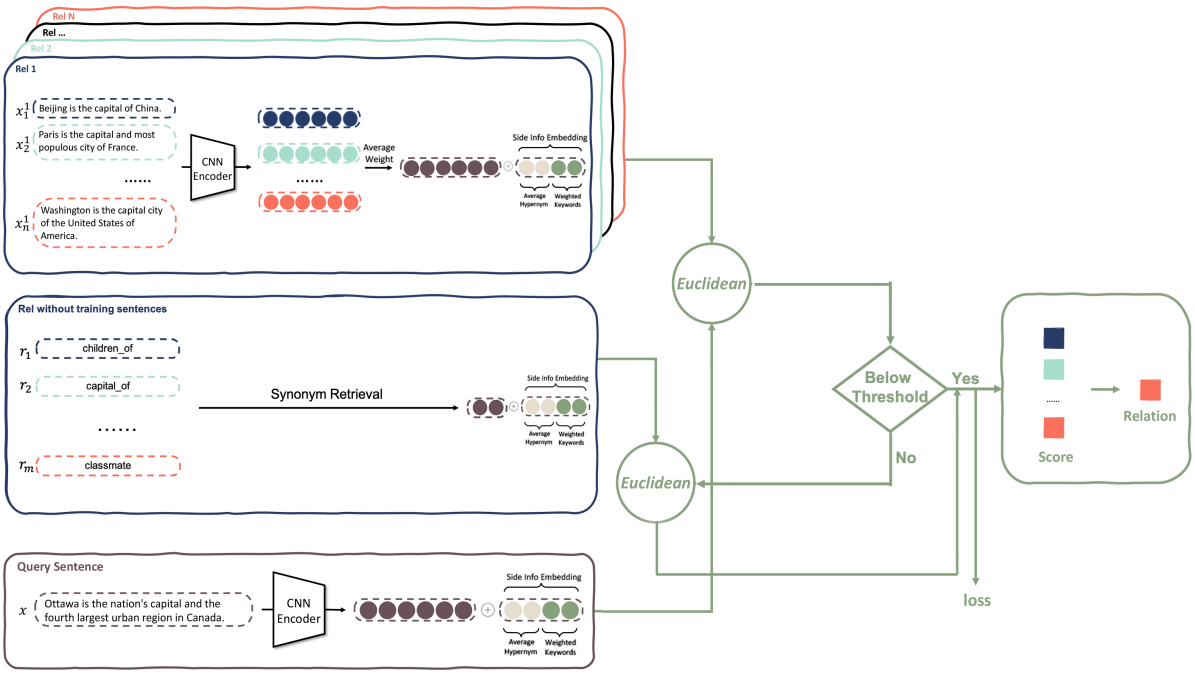


Figure 2.2: Model of Zero-shot Learning for Relation Classification (ZSLRC)

### 2.3.1 Sentence Encoder

The inputs of the ZSLRC model are a set of sentences $\{x_1, x_2, x_3, \cdots x_n\}$ and its corresponding entity pair. For relations with training sentences, our model measures the probability of each relation $r'$ by measuring the distance between query sentences and the average weight of training sentence embeddings. For relations without training sentences, the probability of $r'$ is done by measuring the distance between side information from query sentences and side information from relation types.

#### Word Embeddings

Word embeddings aim to map words or phrases from vocabulary to vectors of numerical forms. The distributed representations are learned based on the usage of words, which allows words that are used in similar ways to result in having similar representations, naturally capturing syntactic and semantic meanings of the words. In this chapter, we first tokenize and lemmatize all words in a sentence, and a 50-dimension GloVe, a pre-trained global log-bilinear regression model for the unsupervised learning of word representations, is used as our initial word embeddings [155]. If the words are out of vocabulary, they are randomly

embedded first, and the vectors are updated while the model is training. Word embedding vectors are updated through training the model.

### Position Embeddings

Word positions also play an essential role in relation classification. Words closer to name entities have more influence on the determination of relation types. We use position features, a combination of relative distances from the current word to both entities, to identify entity pairs [244]. After concatenating position embeddings and word embeddings, the vector representation transforms a sentence into a matrix $S \in \mathbb{R}^{s \times d}$, where s is the sentence length and $d = d_w + d_p \times 2$. For each word $w \in S = \{w_1, w_2, \cdots w_n\}$, its embedding $\hat{w}_i$ is initialized as follows:

$$\hat{w}_i = w_i \oplus p_{i1} \oplus p_{i2} \tag{2.1}$$

where $w_i$ is the pre-trained word vector and $p_{i1}$, $p_{i2}$ are two corresponding position embeddings of the current word with two name entities. Symbol $\oplus$ indicates the concatenation operator. The matrix $S$ is then fed into the CNN encoder.

### CNN Encoder

Because convolutional neural networks can merge all local features and perform the prediction globally, we choose CNN to encode our input embeddings. We learn the instance embedding as follows:

$$x_i = CNN(w_{i-\frac{n-1}{2}}, \cdots, w_{i+\frac{n-1}{2}}) \tag{2.2}$$

$$\hat{x_i} = max(0, x_i) \tag{2.3}$$

$$[s]_j = max\left\{[\hat{x_1}]_j, \cdots, [\hat{x_n}]_j\right\} \tag{2.4}$$

where $CNN(\cdot)$ is a convolutional layer with window size $n$ over the word sequence. A non-linear activation function ReLU is added after the convolutional layer. Function max denotes max-pooling and $[\cdot]_j$ is the j-th value of a vector.

Figure 2.3 shows the architecture of the CNN encoder used in this chapter. Due to time complexity, We simply use one convolutional layer, one non-linear layer, and one max pooling layer to get the sentence embedding. The parameter settings are described in Section 2.4.3

### Side Information Embeddings

Label information and keywords in each sentence also play an essential role in improving the performance of relation classification. For relations without any training sentences, hypernyms, labels, and their corresponding synonyms are used as side information. Side information embeddings are concatenated to the prototype for each relation after the CNN encoder.

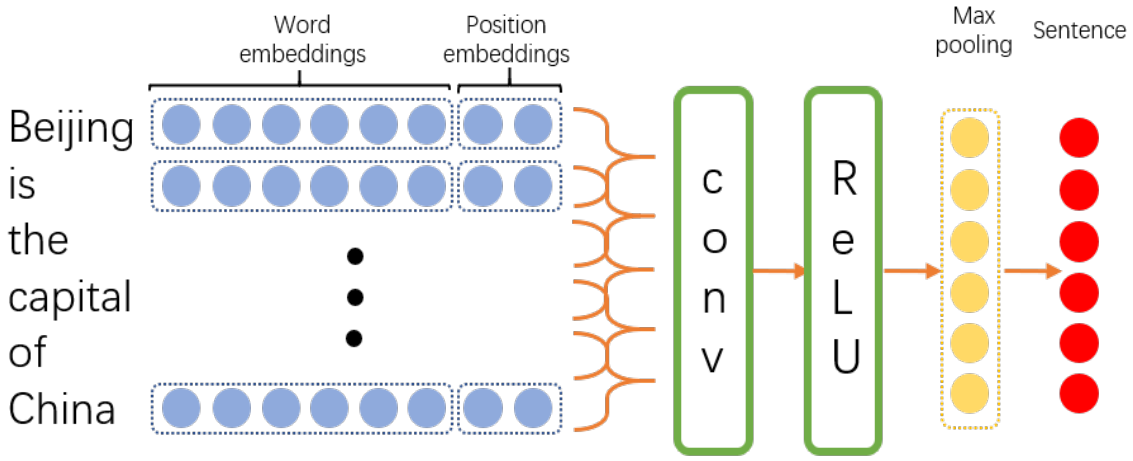


Figure 2.3: CNN Encoder

The final prototype including side information for each relation can be expressed as follows:

$$c_i{}' = \begin{cases} r \oplus si_h \oplus si_r \oplus si_k & r \neq 0 \\ si_h \oplus si_r \oplus si_s & r = 0 \end{cases} \tag{2.5}$$

where $r$ is the initial prototype for each relation, $si_h$ represents the side information from hypernyms, $si_r$ is the side information from relation types, $si_k$ is the side information from keywords in all training sentences of one relation type and $si_s$ is the synonyms for relation types. Details for side information description and its extraction will be described in Section 2.3.2.

### 2.3.2 Side Information Extraction

Side information is the auxiliary information used to detect new relation types. For relations with training sentences, side information is the hypernyms of two entities, the relationship between two entities, and keywords from all training sentences with the same relation type. For relations without training sentences, the side information is hypernyms of two entities by manually labeling, the relation types themselves, and synonyms of the relation types. For query sentences, the side information is hypernyms of two entities and keywords extracted from the sentence.

In this section, we describe hypernyms extraction and keyword extraction in detail because the relationship can be easily obtained from labels, and synonyms of relation types can also be easily acquired through WordNet or other dictionaries [142].

### Hypernyms Extraction

A hypernym is the broad meaning of more specific words. For example, an animal is a hypernym of a dog. The hypernym of two entities in one sentence is extremely important for relation classification. Figure 2.4 shows an example of different sentences with different hypernyms, indicating that hypernyms can help classify different relation types. For example, relation *capital_of* can only occur between two locations, and relation *child_of* can only occur between two people.

| Sentence | Hypernym (Entity) 1 | Hypernym (Entity) 2 |
|---|---|---|
| Beijing is the **capital of** China. | Location (Beijing) | Location (China) |
| The **capital of** France is Paris. | Location (France) | Location (Paris) |
| Mayday Parker is the **child of** Peter Parker. | Person (Mayday) | Person (Peter) |

Figure 2.4: Example of sentences with different hypernyms.

Hypernyms of entities are not easy to acquire. Some existing tools, such as WordNet can only acquire hypernyms from limited vocabularies. In our experiments, less than 10% of entities can achieve their corresponding hypernyms through WordNet. Some previous works used entity types (hypernyms) defined by FIGER as side information [119, 200]. However, only 112 entity types are provided by FIGER, and only 38 of them are used as entity types by [200]. Most of the name entities from sentences in the real world can not get their hypernyms based on this list due to its fixed size and limited entity types. Therefore, we provide an approach for extracting hypernyms through external help from the web.

Hypernyms can be discovered through the definition of entities. We build an automatic hypernym extraction framework based on WordNet, Merriam Webster [1] and Wikidata [2]. Merriam-Webster includes a part of speech description to distinguish nouns of a person (biographical) from nouns of location (geographical). In the real world, there are quite a number of relations that occur between these two hypernyms. Wikidata provides definitions for different entities. We crawl the definition for each entity through Wikidata and get the first Noun as hypernym. For example, *Jeff Bezos is the founder of Amazon.* Commerce is extracted as a hypernym for Amazon. Most entities, including person, location, or other nouns, can get their hypernyms through our proposed framework. The entire framework of hypernym extraction is described in detail in Algorithm 1.

[1] https://www.merriam-webster.com/
[2] https://www.wikidata.org/

**Algorithm 1** Hypernym Extraction

**Input** : sentences $\{x_1, x_2, x_3, \cdots x_n\}$ with same relation.
**Output** : hypernyms of two entities from one relation.
Step 1: Initialize hypernyms to *none*.
Step 2: Find hypernyms $\left\{h_1^1, h_1^2, \cdots h_1^n\right\}$ and $\left\{h_2^1, h_2^2, \cdots h_2^n\right\}$ of entities from WordNet.
Step 3: $h_1 = major\left\{h_1^1, \cdots h_1^n\right\}$, $h_2 = major\left\{h_2^1, \cdots h_2^n\right\}$.
**if** $h == none$ **then**
  go to Step 4.
**else**
  End
Step 4: Getting PoS descriptions $PD$ of entities $E = \left\{e_1^1, \cdots e_1^n\right\}$ and $\left\{e_2^1, \cdots e_2^n\right\}$ from Merriam Webster. $h = Tokenize(PD)$
**if** $h == none$ **then**
  go to Step 5.
**else**
  End
Step 5: Crawling definitions $D$ for $E$ from Wikidata. $h$ = first Noun of $Tokenize(D)$.

**Keywords Extraction**

The keyword is another crucial factor of side information because it reflects the importance of the featured item. TF-IDF (term frequency-inverse document frequency) is used for keyword extraction due to its efficiency [167]. It estimates the frequency of a word in one sentence over the maximum in a collection of sentences with the same relation type and assesses the importance of a word in one set of sentences. For relations with training sentences, all sentences are aggregated as one document $d$, and TF-IDF is implemented based on the document. Other models can also be used for keyword extraction.

### 2.3.3 Prototypical Network with Side Information Embedding

Instead of adding a softmax layer directly after encoders for relation classification, we use prototypical networks to compute a prototype for each relation after encoders because some works show that prototypical networks work well for few-shot learning [51, 190]. They are simpler and more efficient than other meta-learning algorithms, making them suitable for few-shot or zero-shot learning tasks. By comparing the distance between query sentences with prototypes for each relation, we can classify the relation. In this section, we describe the prototypical network model and its transformation with weighted side information embedding for zero-shot learning to detect new relations.

The main idea for the prototypical network is to compute a prototype representing each

relation. Each prototype is the mean vector of embedded sentences belonging to one relation.

$$c_i = \frac{1}{N}\sum_{i=1}^{N} f_\phi(x_i) \tag{2.6}$$

where $c_i$ represents the prototype for each relation $r_i$ and $f_\phi$ is an embedding function, which is a CNN encoder in our model. Instead of concatenating all hypernyms and keywords directly after each prototype, we argue that not all keywords are of equal importance. To determine a more accurate representation for each relation, we calculate a weighted side information embedding for each relation. The equation of side information embedding $si$ is as follows:

$$si = f(\frac{h_1 + h_2}{2}) \oplus f(k_1) \oplus \cdots \oplus f(k_n) \oplus K \tag{2.7}$$

$$K = \sum_{m-n}^{m} (\frac{\alpha_i}{\sum_{i=m-n}^{m} \alpha_i} f(k_i)) \tag{2.8}$$

where $f(\cdot)$ is a word embedding model, $h_1$ and $h_2$ are two hypernyms for name entities and $k_i$ denotes the keyword. Symbol $\oplus$ is the concatenation operator, n is determined by exploration search, $m$ is the total number of keywords and $\alpha_i$ is a calculated weight by:

$$\alpha_i = \frac{count(k, s)}{size(s)} \cdot log(\frac{N}{sentence(k, S)}) \tag{2.9}$$

where $s$ is each instance and $N$ is the number of instances in a relation. The final representation for each prototype with side information embedding $ps_i$ can be expressed by:

$$ps_i = c_i \oplus si_i \tag{2.10}$$

The probabilities of the relations in $\Re$ for a query instance $x$ is computed as follows:

$$p_\phi(y = ps_i|x) = \frac{exp(-d(f_\phi(x), ps_i))}{\sum_{ps_i'} exp(-d(f_\phi(x), ps_i'))} \tag{2.11}$$

where $d(.)$ is the Euclidean distance function as below:

$$d(f_\phi(x), ps_i) = \sqrt{\sum_{i=1}^{n} (ps_i - f_\phi(x))^2} \tag{2.12}$$

We use Euclidean distance instead of cosine similarity for distance calculation because previous work shows that Euclidean distance can improve performance substantially over cosine similarity [190]. We have not added any attention layer in our final model because (1) previous work shows there is little improvement in performance compared with vanilla prototypical networks [51]; (2) Ablation study in Section 2.4.4 shows there is no improvement on ZSLRC with attention layers.

For the zero-shot learning task, each relation is given the embedding for side information of the relation rather than a small number of labeled training sentences. We take the embedding of side information into a shared space to serve as the prototype for each relation. The core idea in traditional prototypical networks is to use an average embedding to represent a class [51, 190]. If there is no training data in that class, a high-level description of the class is used to represent that class. We modify prototypical networks to deal with both relations with training sentences and relations without training sentences. The difference between traditional prototypical networks and our proposed model is that they calculate the distance between the query sentence and prototype of each class to find the nearest one. Our proposed model first decides whether the query sentence is in a class with training data or one without any training data based on a threshold. The reason is that finding the nearest distance directly based on all classes (with training data and without training data) is not fair for the class without training data because the high-level description is too general that it always has a longer distance compared with the classes that have training data.

**Algorithm 2** Algorithms for New Relation Extraction

**Input** : prototype for each relation $c_i$, testing sentence $x$, threshold $t$.
**Output** : relationship $r$ of $x$.
Distance Calculation. $d(f_\phi(x), c_i)$.
Take $v = exp(-d(f_\phi(x), c_i))$.
**if** $v > t$ **then**
    Classification of known relations. $r = argmax(\frac{v}{\sum_{c_i'} v'})$
**else**
    Take side information embedding. $f_\phi(x)[SI\_DIM :]$
    Distance Calculation. $d(f_\phi(x)[SI\_DIM :], c_i)$.
    Take $v_{new} = exp(-d(f_\phi(x)[SI\_DIM :], c_i))$.
    Softmax of new relations. $r_{new} = argmax(\frac{v_{new}}{\sum_{c_i'} v'_{new}})$

We modify the prototypical network as follows: We first compare the distance between an input sentence with each prototype of known relations. The key mechanism for extracting new relations is that if the above distance is larger than a threshold, we consider the sentence to have a new relation. Then we take the side information embedding of the input sentence and compare the distance between it with prototypes for new relations. Then we use a softmax layer to compute the probabilities for each new relation. The threshold selection is essential because it influences the decision of a relation type as an existing relation or a new relation. We implement a grid search to select the optimal threshold on the validation set. The entire framework of the ZSLRC model to deal with a combination of known relations and new relations is described in Algorithm 2.

## 2.4 Experiments

In this section, we conduct several experiments on two public datasets: NYT [173] and FewRel [72] to show that our proposed model outperforms other existing models on both a noisy dataset with a large number of training sentences and a clean dataset with few training sentences. We design experiments for generalized zero-shot learning tasks and provide a detailed analysis to show the effectiveness and advantages of our proposed model.

### 2.4.1 Datasets and Evaluation Metrics

In our experiments, we evaluate our model over two widely used datasets: the NYT dataset [173] and the FewRel [72] dataset. In the following, we describe each dataset in detail.

- ***NYT [173].*** The NYT dataset was generated by aligning Freebase relations with the New York Times corpus (NYT). There are 53 possible relationships in total. It is an unbalanced noisy dataset because all the relationships have a different number of sentences.
- ***FewRel [72].*** The FewRel dataset is a human-annotated few-shot RC dataset consisting of 80 types of relations, each of which has 700 instances.

To fairly compare the performance of our proposed model with other state-of-the-art models in supervised learning and few-shot learning tasks, we use the same training, validation, and testing set of the NYT dataset and the same training and validation set of FewRel. We evaluate our proposed model on the validation set of FewRel because the test set is not available directly. In order to properly evaluate the performance of our proposed model in a zero-shot learning task, we re-split the above two public datasets for training, validation and testing set. Details of dataset re-splitting and experiment design are introduced in section 2.4.2. Note that we do not use any other clean, supervised dataset such as SemEval-2010 Task 8 (SemEval) because this dataset only contains 19 kinds of relations, which is less persuasive when re-splitting the dataset to evaluate the performance of our proposed model in zero-shot learning task [76].

The evaluation metrics adopted in this chapter are the standard micro Accuracy (Acc.), Precision (Prec.), Recall (Rec.), and F1-score, similar to those used for the baseline.

### 2.4.2 Experiment Design

In a real-world scenario, there exist both kinds of relations with training instances and without any training instances. To make it simple and clear to understand, we call the relations

with training instances known relations and the relations without any training instances new relations in the following discussion. To evaluate the effectiveness and robustness of our proposed model in a zero-shot learning task, we design the testing cases to contain different percentages (from 0% to 100% with a step of 10%) of new relations. Note that 0% means a thoroughly supervised learning or few-shot learning scenario, whereas 100% means a completely zero-shot learning scenario. The experiment design for zero-shot learning relation classification follows the criteria of zero-shot text classification; the different rates of unseen classes are used in testing cases [247].

***NYT [173].*** NYT is an unbalanced noisy dataset with 53 different relationships in total. We added initial training, validation, and testing sets together and re-split the dataset into ten types of relations for the training pool. Each relation has over 10k sentences, and the rest relations are for the validation pool and testing pool. In the training pool, we take 10k sentences of each relationship for training, and the rest types of relations are used to validate and test known relations. In all, we have 100k sentences of 10 relationships in total for training, 13k sentences of known relations, and 5k sentences of new relations for validation and testing. For example, if a new relation *capital_of* is allocated to the testing set, no *capital_of* sentences appear in the training set.

***FewRel [72].*** FewRel dataset has 80 types of relations with 700 instances each. We re-split the dataset into 40 types of relations for training and 40 types of relations for testing. There are no overlapping relations among the training and testing sets. To evaluate our proposed model in a real-world scenario (a combination of known and new relations in the testing set), we take 300 instances from each relation type in the training set to make a testing pool containing known relations. In total, we have 40 relations, and each relation has 400 instances in the training pool, 40 known relations. Each relation has 300 instances in the testing pool, 40 new relations, and each relation has 700 instances in the testing pool.

### 2.4.3 Parameter Settings

For all the models, we use the pre-trained word embeddings with a 50-dimensional Glove model (6B tokens, 400K vocabulary) and a randomly initialized 5-dimensional position embedding on NYT corpus for initialization [155]. Both word embeddings and position embeddings are trainable during training. The number of feature maps in the convolutional layer is 800, and the side information embedding dimension is 300. We experimentally study the effects of two crucial parameters on our model, learning rate $\alpha$ and threshold $t$. We use a grid search to select the optimal learning rate $\alpha$ for SGD among $\{1e-1, 1e-2, 1e-3, 1e-4\}$ for minimizing the loss, the threshold $t$ for determining a new relation among $\{2e-08, 7e-08, 2e-07, 7e-07\}$ on a validation set with 20% of new relations. The range of threshold is determined by the minimum and maximum values of $e^{-d}$ on a validation set, where $d$ is the Euclidean distance between the query sentence and prototype for each relation. For other parameters, we follow the settings used in previous works so that

Table 2.1: Parameter Settings

| Parameter | Value |
|---|---|
| Word Embedding Dimension $d_w$ | 50 |
| Position Embedding Dimension $d_p$ | 5 |
| Side Information Embedding Dimension $d_{si}$ | 300 |
| Hidden Layer Dimension $d_h$ | 800 |
| Convolutional Window Size $n$ | 3 |
| Batch Size | 1 |
| Initial Learning Rate $\alpha$ | 0.01 |
| Weight Decay | $10^{-5}$ |
| Threshold $t$ | 2e-08 |

our model can be fairly compared with these models [51, 244]. Table 2.1 shows parameters used in our experiment.

### 2.4.4 Results

#### Baseline Methods

We compare our proposed model to several state-of-the-art models in both supervised learning and few-shot learning tasks. For a supervised learning task on the NYT dataset, we compare our model with **CDNN**, which first proposed the idea of position embedding [244]. The reason we chose this model to make the comparison is that we both use similar CNN encoders so the improved performance of our model is not because of using any better encoders such as BERT [34]. The reported result for **CDNN** is our re-implementation on NYT because the source code is not available, and their original report is the evaluation on other datasets [244]. The reported result for the **REDN** is from the original published literature [105]. Note that **REDN** is a relation classification model using the given name entities, and we only copy the result of the single relation classification of this chapter so that we can make a fair comparison. For the few-shot learning task on the FewRel dataset, we compare our model with **Meta Network**, **GNN**, **SNAIL**, **Proto**, **Proto-HATT** and **Proto-CATT(CNN)**. The six baselines above on the FewRel dataset are reported by [87], which are all current state-of-the-art FSL models. Note that the above FSL model Proto-HATT and our proposed model use the same pre-trained word embedding model 50-dimension GloVe, CNN encoders, and the same training parameters only except batch size and hidden layer dimension. For zero-shot learning, we compare our proposed model with the re-implemented **CDNN**, **REDN**, **Proto** and **Proto-HATT** on our re-split NYT and FewRel datasets to show the effectiveness and robustness of our proposed model.

Table 2.2: Results of different models on NYT (%). Our re-implementation is marked by $*$.

| Model | Precision | Recall | F1 |
|---|---|---|---|
| CDNN* [244] | 46.4 | 52.7 | 45.8 |
| REDN [105] | 95.1 | 94.0 | 94.6 |
| ZSLRC | **98.1** | **97.9** | **97.6** |

**Results on NYT**

Table 2.2 demonstrates that our proposed model achieves a substantial gain in precision, recall and F1-score over other baselines for the supervised learning task. We compare the ZSLRC model with CDNN [244] as both models use a CNN encoder. The results show that ZSLRC achieves a significant performance improvement in precision, recall, and F1-score. Our proposed ZSLRC also outperforms a recently proposed method (REDN) [105] by **3%** precision, **3.9%** recall and **3%** F1-score though REDN uses BERT encoder. This is important to note because BERT-based sentence encoders have significantly outperformed other sentence encoders including our proposed one-layer CNN-based type [87].

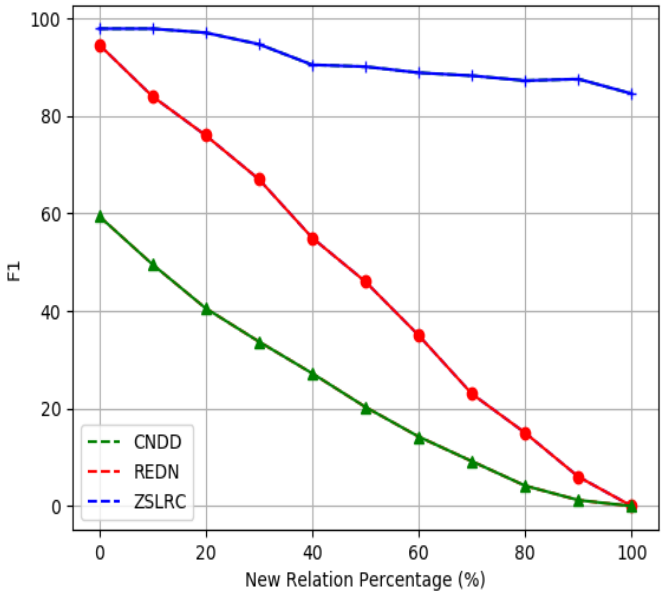


Figure 2.5: F1-score of ZSLRC when different proportions of new relations appear in the NYT dataset.

The achieved performance improvement indicates that the proposed side information is competitively beneficial for relation classification. To evaluate our proposed model in a real-world scenario, we re-split the NYT dataset and use 40+ relations as new relations with no labeled

Table 2.3: Ablation Results on NYT dataset (Accuracy%)

| | 10% | 30% | 50% | 70% | 90% |
|---|---|---|---|---|---|
| ZSLRC(HE) | 88.94 | 70.57 | 52.12 | 33.87 | 15.48 |
| ZSLRC(KE) | 93.12 | 82.22 | 71.00 | 60.47 | 49.07 |
| ZSLRC(SIE) | 93.86 | 85.14 | 81.91 | 78.79 | 72.57 |
| ZSLRC(WSIE) | **96.64** | **94.46** | **92.14** | **91.82** | **89.3** |

training data. As is shown in Figure 2.5, 0% of new relations means it is a supervised learning task and all relations in the testing set have corresponding labeled training data. 100% of new relations means it is a conventional zero-shot learning task, and all relations in the testing set do not have any labeled training data. We compare the performance of our proposed model with CDNN [244] and REDN [105] as we vary the percentage of new relations in the testing set. As shown in Figure 2.5, the F1-score of both CDNN and REDN decreases when the percentage of new relations increases. This is because the model can not detect new relations and instead classifies the new relation as one of the existing relations in the training set. That is why the F1-score becomes zero when the new relation percentage is 100%. The F1-score of our proposed model ZSLRC only drops around 15% from a fully supervised case to a zero-shot case, indicating that our model is effective and sufficiently robust when dealing with new relations.

To investigate the contribution of different side information embeddings in ZSLRC, we conduct an ablation study in zero-shot learning settings by adding each component, including hypernyms embedding (HE), keywords embedding(KE), side information embedding(SIE) and weighted side information embedding(WSIE). Table 2.3 shows the results of the ablation study different proportions of new relations in the testing set. We find out that all kinds of side information embedding help detect new relations. Only adding hypernyms embedding to the model can help detect new relation classes. However, the accuracy rate drops significantly from 88.94% in 10% of new relations to 15.48% in 90% of new relations. This is because hypernyms only represent the main categories for name entities and could help classify the relations roughly without training instances. Compared with hypernyms embedding, keywords embedding achieves much better performance because keywords (keywords extracted from training instances of seen class and synonyms of labels of unseen class) represent discriminative features of each instance, shortening the distance between query instance and prototype. Nevertheless, the performance of ZSLRC(KE) still drops considerably when the percentage of new relations increases. ZSLRC(SIE) achieves a significant accuracy performance improvement. Side information embedding is a combination of hypernyms embedding and keywords embedding. It represents high-level information about the instance, shortening the distance of instances with the same relation. As shown in Table 2.3, it is more robust when the percentage of the new relation class increases. Since we assume that not all side information is of equal importance, we also implement ZSLRC with weighted side information added to the model as introduced in Section 2.3.3. This model achieves the best

Table 2.4: Results of Accuracy Comparison Among Models (%)

| Model | 5-w-1-s | 5-w-5-s | 5-w-10-s | 10-w-1-s | 10-w-5-s | 10-w-10-s |
|---|---|---|---|---|---|---|
| Meta Network* | 64.46 | 80.57 | - | 53.96 | 69.23 | - |
| GNN* | 66.23 | 81.28 | - | 46.27 | 64.02 | - |
| SNAIL* | 67.29 | 79.40 | - | 53.28 | 68.33 | - |
| Proto(CNN) | 73.62 | 85.78 | 88.45 | 60.96 | 75.38 | 78.71 |
| Proto-HATT(CNN) | 74.68 | 86.73 | 89.64 | 61.61 | 77.04 | 79.99 |
| Proto-CATT(CNN) | - | 87.48 | 89.28 | - | 77.46 | 80.39 |
| **ZSLRC(CNN)** | **75.83** | **87.84** | **89.67** | **63.54** | **77.64** | **80.69** |

Note that to fairly compare the performance of each model, we only compare the models with the same 50-dimension GloVe embedding and CNN encoders of the same parameters. Better results can be achieved through the BERT encoder.

performance. Besides the high accuracy performance with any proportions of new relations, it is also robust enough that it only drops 7.3% accuracy rate from 10% of new relations to 90% of new relations. When the proportions of new relation increase, the accuracy of ZSLRC with weighted side information embedding drops less than the other models.

**Results on FewRel**

The evaluation results of few-shot learning on FewRel are shown in Table 2.4. Note that results with * are reported in [72]. The result of the Proto-CATT model is copied from their original paper because of no public code [87]. We re-implement Proto and Proto-HATT with all parameters the same except the hidden layer dimension. Both Proto-HATT and Proto-CATT use CNN encoders and attention layers to help improve the performance. To fairly compare the effectiveness of side information embedding, we only compare our models with other state-of-the-art models using CNN encoders with attention layers. Each task is provided with a set of k-labeled sentences from each of the N classes that have not previously been trained upon. We conduct the experiments of N-way K-shot few-shot learning tasks following the method introduced in [153]. Table 2.4 shows that ZSLRC (without any attention layer) outperforms the other state-of-the-art models using multiple attention layers on several N-way K-shot tasks, especially for 1-shot cases. The accuracy of our proposed model on 5-way 1-shot and 10-way 1-shot tasks are 75.83% and 63.54%, which is 1.15% higher and 1.93% higher than the model Proto-HATT. Next, we investigate ZSLRC performance on N-way one-shot learning. Figure 2.6 demonstrates changes in accuracy as the number of ways changes in comparison with two state-of-the-art models. As the number of classes increases, the accuracy drops, but our proposed model has a slower dropping rate than other models. We conjecture that both the increased difficulty of a larger number of ways and the side information embedding we have proposed enable the ZSLRC to make more fine-grained decisions and is therefore more robust to the increased complexity introduced by

more classes.

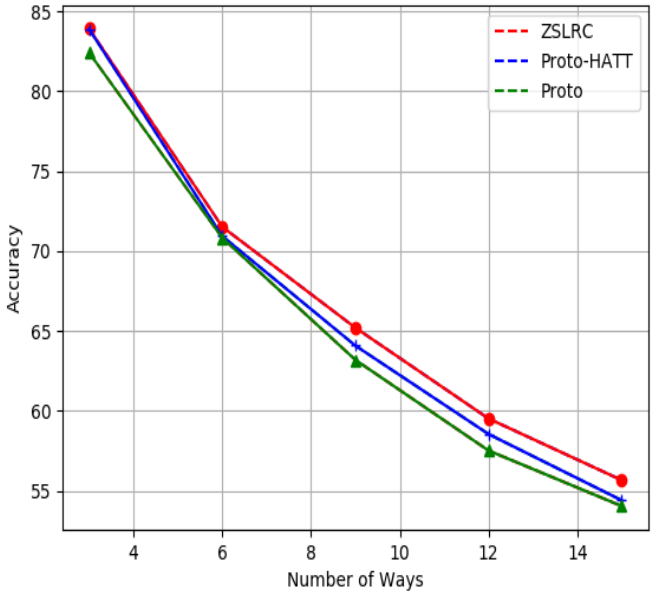


Figure 2.6: Accuracy of our proposed model in different N-way One-shot tasks.

To evaluate the effectiveness and robustness of ZSLRC in a generalized zero-shot learning task, we evaluate our models on the re-split FewRel dataset. To test the effectiveness and robustness of our proposed model, we compare our proposed model ZSLRC with Proto(CNN) and Proto-HATT(CNN) [51] in zero-shot settings described in Section 2.4.2. Figure 2.7 shows the performance of ZSLRC in a real-world scenario with different percentages of new relations on the re-split FewRel dataset. The accuracy of ZSLRC only drops from 97.3% to 86.8%, indicating the effectiveness and robustness of our proposed model for recognizing new relations in the real world. We show that zero-shot learning to new relation types is possible and we set the bar for future work on this task.

We also conduct an ablation study on the FewRel dataset to learn the effectiveness of weighted side information embedding. Besides the models introduced in Section 2.4.4, we also implement a new model with attention layers for weighted distance (WSIEA), to investigate the influence of the attention layer. Table 2.5 shows the results of the ablation study. We can observe that all kinds of side information embedding contribute to the performance of ZSLRC. There is a big accuracy performance improvement when hypernyms embedding introduced in Section 2.3.2 is added to the model because hypernyms represent a general embedding for different name entities, which will decrease the variance from different word embeddings, leading to a shorter distance. Keyword embedding also contributes significantly to the performance, indicating the importance of keywords to side information embedding. Similar to the ablation result on the NYT dataset as shown in Section 2.4.4, using side

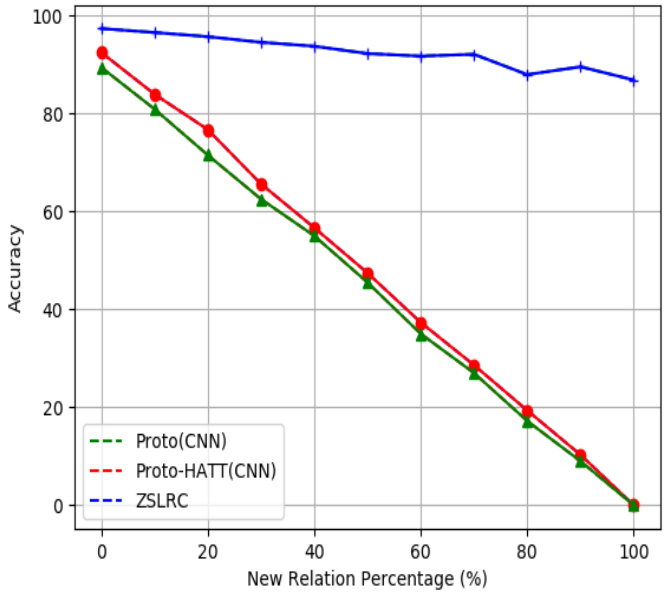


Figure 2.7: Accuracy of ZSLRC when different proportions of new relations appear in re-split FewRel dataset.

information embedding helps improve the performance, and the model with weighted side information embedding achieves the best performance. We also added an attention layer built by three neural network layers and a softmax layer on top of each prototype to calculate linear separability based on the distribution of each prototype's sentence representations. However, there is no improvement in the attention layer. We conjecture that the weighted side information embedding has already captured each relation's vital feature. In this way, merely using side information embedding helps simplify the model's architecture, reducing the complexity of several neural network layers by attention mechanism.

Table 2.5: Ablation Results on FewRel dataset (%).

| Model | 5-w-1-s | 5-w-5-s | 5-w-10-s | 10-w-1-s | 10-w-5-s | 10-w-10-s |
|---|---|---|---|---|---|---|
| Proto(CNN) | 73.62 | 85.57 | 88.17 | 62.22 | 75.01 | 78.50 |
| ZSLRC(HE) | 75.66 | 86.55 | 88.98 | 63.28 | 76.58 | 79.93 |
| ZSLRC(KE) | 74.57 | 86.70 | 89.09 | 62.39 | 76.99 | 80.06 |
| ZSLRC(SIE) | 75.56 | 87.34 | 89.17 | 63.02 | 77.16 | 80.34 |
| **ZSLRC(WSIE)** | **75.83** | **87.84** | **89.67** | **63.54** | **77.64** | **80.69** |
| ZSLRC(WSIEA) | 75.58 | 87.16 | 89.17 | 62.85 | 76.71 | 80.18 |

## 2.5 Summary

We propose ZSLRC[3], a zero-shot learning relation classification framework based on modified prototypical networks. ZSLRC can detect new relations with no corresponding labeled data available for training. ZSLRC utilizes weighted side information constructed from labels, keywords, and hypernyms of entities extracted from our proposed automatic hypernym extraction framework. We evaluate our model on supervised learning, few-shot learning, and zero-shot learning tasks. The results demonstrate that our proposed ZSLRC outperforms other state-of-the-art models in all tasks. In addition, the results demonstrate the effectiveness and robustness of our proposed model. In future work, we plan to explore the following directions: (1) Due to the surprising performance improvement contributed by side information embedding, we will explore different ways to embed side information, leading to learning different representations of each prototype (relation). (2) We will explore using other popular sentence encoders such as BERT to improve the performance for relation classification.

[3]Implementation details can be accessed via: https://github.com/gjiaying/ZSLRC

# Chapter 3

# Prompt-based Zero-shot Relation Triplet Extraction with Semantic Knowledge Augmentation

In relation triplet extraction (RTE), recognizing unseen relations for which there are no training instances is a challenging task. Efforts have been made to recognize unseen relations based on question-answering models or relation descriptions. However, these approaches miss the semantic information about connections between seen and unseen relations. In this paper, We propose a prompt-based model with semantic knowledge augmentation (ZS-SKA) to recognize unseen relations under the zero-shot setting. We present a new word-level analogy-based sentence translation rule and generate augmented instances with unseen relations from instances with seen relations using that new rule. We design prompts with weighted virtual label construction based on an external knowledge graph to integrate semantic knowledge information learned from seen relations. Instead of using the actual label sets in the prompt template, we construct weighted virtual label words. We learn the representations of both seen and unseen relations with augmented instances and prompts. We then calculate the distance between the generated representations using prototypical networks to predict unseen relations. Extensive experiments conducted on three public datasets FewRel, Wiki-ZSL, and NYT, show that ZS-SKA outperforms other methods under zero-shot setting. Results also demonstrate the effectiveness and robustness of ZS-SKA.

## 3.1 Introduction

Relation triplet extraction (RTE) aims to extract both the pairs of entities and relations from unstructured text. However, existing approaches based on supervised learning [86, 104, 136, 170, 262, 264] or few-shot learning [37, 42, 50, 70, 131, 171] still require labeled data. They can not catch up with a dynamic and open environment where new classes emerge. In the real-world setting, the classes of instances are sometimes rare or never seen in the training data. Thus, we tend to learn a model similar to the way humans learn and recognize new concepts. Such a task is referred to as zero-shot learning (ZSL). We follow the same definition of ZSL in [16, 29] to conduct experiments.

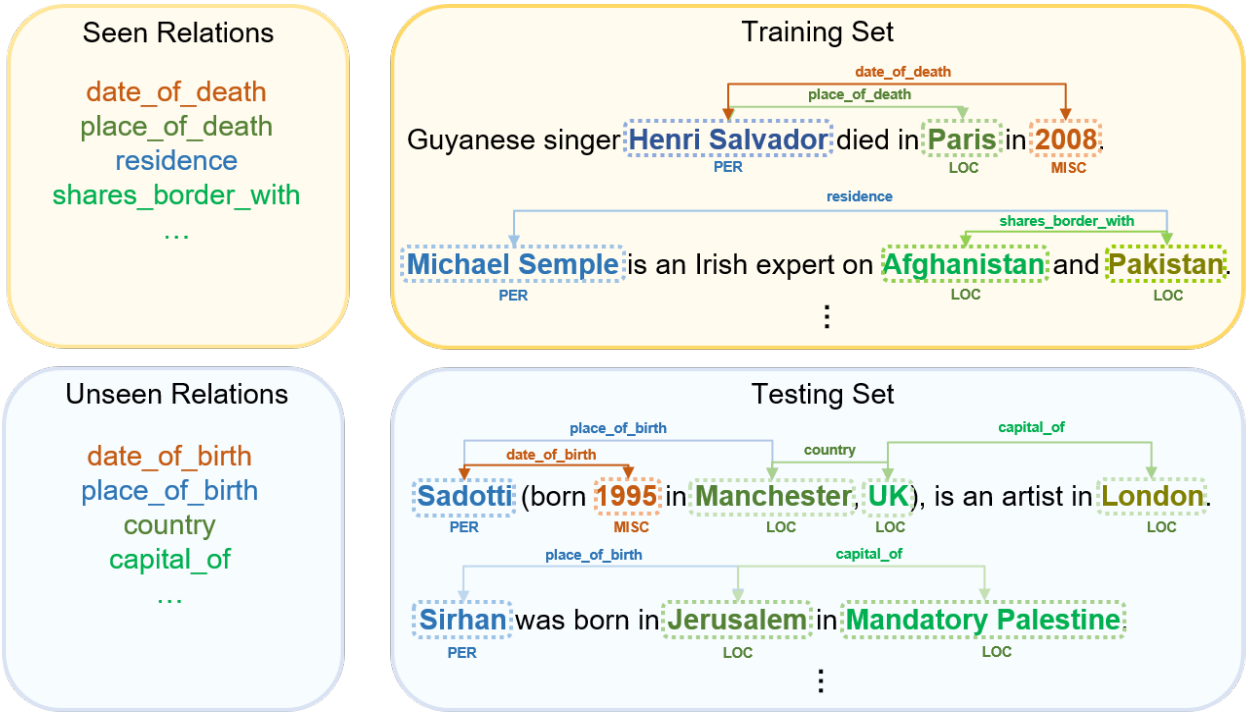


Figure 3.1: Zero-shot RTE. There is no overlap of classes between training and testing data.

Zero-shot RTE aims to extract relation triplets in a sentence that is absent from the learning stage. Figure 3.1 shows an example of zero-shot RTE. The relation sets at the training and testing stages are disjoint. The model for zero-shot RTE is only trained on the seen relations in the training stage and extracts triplets with unseen relations in the testing stage. Existing approaches to zero-shot relation extraction still have limitations. First, some models perform zero-shot relation extraction by question answering [101] or by using GPT-2 to help generate synthetic data [29]. These models have a strong assumption that an excellent additional deep learning model is learned and that all values extracted from this model are correct. Second, some existing studies formulate relation extraction as a text entailment task [154]. They only predict a binary label indicating whether the name entities in the given sentence can be described by a given description. Third, some SOTA models leverage auxiliary information to tackle zero-shot tasks. They focus on class names/descriptions, losing the connection or relationships between seen relations and unseen relations [16, 61, 208]. Besides, these works mainly focus on zero-shot relation classification (ZSRC), which only predicts unseen relations instead of triplets in the format of <head entity, relation, tail entity>. ZSRC has a strong assumption that two name entities are available for training. However, it is not realistic that name entities are already provided.

To address the above challenges, we propose a prompt-based model with semantic knowledge augmentation (ZS-SKA) to perform zero-shot RTE. We first implement data augmentation by a word-level sentence translation to generate augmented instances with unseen relations from training instances with seen relations. We follow a new generation rule introduced in Sec. 3.3.2 to generate high-quality augmented instances for training in zero-shot settings. Note that ZS-SKA is trained only on labeled data from seen classes and augmented data generated from seen classes.

Secondly, inspired by prompt-tuning on pre-trained language models [182, 183], we design the prompts based on the knowledge graph to integrate semantic knowledge to generally infer the features of unseen relations using patterns learned from seen relations. For the prompt design, we consider semantic knowledge information, including relation descriptions, super-class of relations and name entities, and a general knowledge graph to effectively learn the unseen relations. Instead of using the real label word directly in the prompt template, we automatically search a set of appropriate label words based on the knowledge graph for each label. The weight of each appropriate label word is calculated based on its semantic knowledge information in Sec. 3.3.2. We calculate the distance between each appropriate label with the true label itself to help denoise the set of appropriate label words. Then, we construct virtual label words in the prompt by weighted averaging all appropriate label word candidates.

Finally, we apply prototypical networks [191] to compute a prototype representing each relation. Each prototype is the mean vector of embedded and augmented sentences with prompts belonging to one relation. Euclidean distance is calculated between query sentence embeddings with prototypes to predict relations. A distance threshold explored in the validation set is applied to determine the number of relation triplets. For name entity predictions,

we use a name entity extractor to recognize different types of entities. Then, sorted entity types are compared with the super-class of name entities in the prompt to determine the final relation triplets. The contributions:

- We propose a prompt-based model with semantic knowledge augmentation (ZS-SKA) to extract triplets with unseen relations under the zero-shot setting. Unlike some previous works, ZS-SKA considers semantic information from different granularities and does not rely on other large models with additional training.
- We present a new word-level sentence translation rule to generate augmented instances with unseen relations from instances with seen relations. The augmented sentences are then used as the training sets for unseen relations.
- We propose prompts for training based on an external knowledge graph to integrate semantic knowledge information learned from seen relations. We construct weighted virtual label words for the mask in the prompt template instead of using the actual label sets.
- We demonstrate that ZS-SKA significantly outperforms state-of-the-art methods for relation extraction with unseen relations under the ZSL setting on three public datasets.

## 3.2 Related Work

### 3.2.1 Prompt Learning in NLP

With the development of Generative Pre-trained Transformer 3 (GPT-3) [11], prompt-based learning has received considerable attention. Language prompts have been proven to be effective in downstream tasks leveraging pre-trained language models [32, 156, 198]. Human-designed prompts have achieved promising results in few-shot learning for sentiment classification [182, 183]. To avoid labor-intensive prompt design, studies explore prompts that are generated automatically [53, 92, 185]. However, most of the studies focus on supervised or few-shot learning on text classification [48, 66, 73, 81], event detection [107], relation classification [24, 73] and name entity recognition [85, 106, 135]. Inspired by these works, we explore prompt-based zero-shot learning in RTE.

### 3.2.2 Zero-shot Relation Classification

Relation classification is the problem of classifying relations given two name entities within a sentence. Most existing works rely on sufficient human-labeled data or noisy labeled data by distant supervision. When no training instances are available, some studies use zero-shot

relation classification to extract unseen relations. This is typically done using question-answering models by listing questions that define the relation's slot values [13, 101]. Some studies formulate relation extraction as a text entailment task [154]. Some studies utilize the accessibility of the relation descriptions to get the information for unseen relations [16, 61, 122, 154, 160, 179]. However, these models only utilize class names semantic information, losing the connections between relations. Other studies focus on establishing the connection between relations with knowledge graph [108] or contrastive learning [208]. Nevertheless, all these works only focus on relation classification instead of relation triplet extraction. In the real world, it is not practical and realistic that two name entities are provided for training. Therefore, we focus on a more complex and realistic task: extracting both name entities and their relations.

### 3.2.3 Zero-shot Relation Triplet Extraction

Current approaches focusing on zero-shot relation triplet extraction (RTE) require large computing resources and additional deep-learning models. For example, RelationPrompt requires to fine-tune a pre-trained GPT-2 (124M parameters) as the relation generator [29]. PCRED needs to train the entity boundary detection module by four neural network layers to get the possible boundaries for name entities in the triplet [99]. ZETT views relation extraction as a template-filling problem, fine-tuning T5 to get the ranking score for potential triplets [96]. However, ZETT can not discriminate against similar relations because the connections of different relations are lost. Inspired by data augmentation from knowledge graph in text classification [20, 247] and prompt-based few-shot learning [81], we propose a prompt-based zero-shot RTE framework (ZS-SKA) incorporating external knowledge from the knowledge graph. Different from these existing works, the data augmentation module and name entity recognition module in ZS-SKA do not require any additional training. ZS-SKA can better catch the connections between relations due to the incorporated knowledge graph in the prompt template.

## 3.3 Methodology

In this section, we introduce the overall framework as shown in Figure 3.2 of ZS-SKA.

### 3.3.1 Problem Definition

To do zero-shot relation extraction, we adopt the problem setting in [29, 61] for zero-shot relation triplet extraction and setting in [16] for zero-shot relation classification. We also conduct ablation experiments following the zero-shot definition in [216] which is a generalized zero-shot setting where partial labels are unseen. Given labeled instances belonging to a set

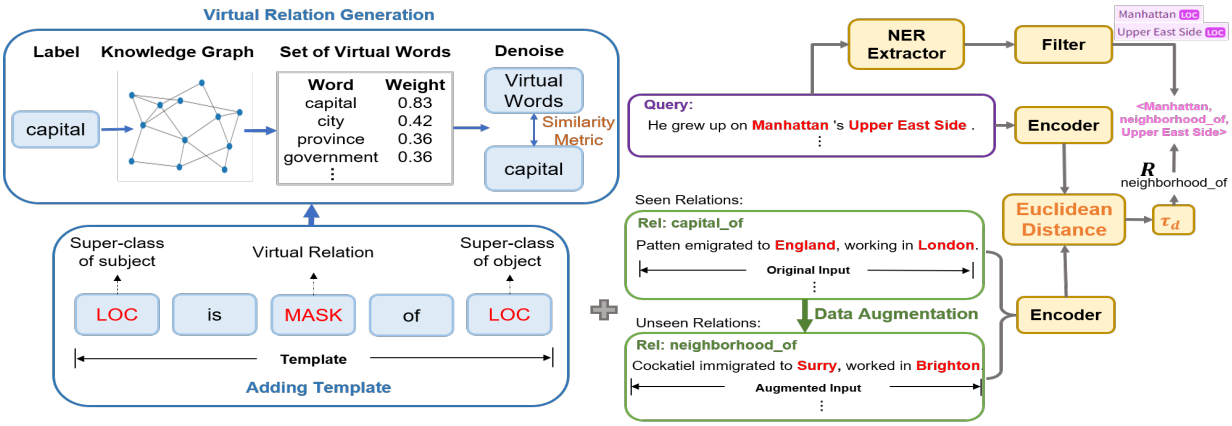


Figure 3.2: ZS-SKA overall architecture with components explained in Sec. 3.3.2.

of seen classes $S$, a model $M : X \rightarrow Y$ is learned, where $Y = Se \cup U$; $U$ is the unseen class. For zero-shot RTE, let $R_s = \{r_s^1, \cdots, r_s^m\}$ and $R_u = \{r_u^1, \cdots, r_u^n\}$ denote the sets of seen and unseen relations, where $m = |R_s|$ and $n = |R_u|$ are the number of relations in the two disjoint sets, i.e., $R_s \cap R_u = \emptyset$. The set of relations $R$ is pre-defined. The input of the training set consists of (1) seen relations $R_s$ with input sentences $X_i$, (2) unseen relations $R_u$ with super-class $S(e_{i_h})$, $S(e_{i_t})$ super-class of two name entities, and (3) external knowledge graph $G$. Here super-class is the hypernym of the item. For example, location (LOC) is the super-class of New York. The output of the model is a triplet in the format of <head ($e_{i_h}$), relation (r), tail ($e_{i_t}$)> or a set of triplets if $X_i$ contains multiple triplets. Our goal is to train a zero-shot relation triplet extraction model $M$ to (1) learn the representations of both seen and unseen relations, (2) predict new triplet $<e_{i_h}, r_u, e_{i_t}>$, where the relation $r_u$ is not seen during the training phase. $M$ is learned by minimizing the semantic distance between the embedding of the input and relation representations built from the knowledge graph $G$. For the zero-shot RC, the only difference with zero-shot RTE is that the two name entities $e_{i_h}$ and $e_{i_t}$ are known information with the input sentence $X_i$. Therefore, the model only needs to predict $r_u$ given $X_i$ with $e_{i_h}$ and $e_{i_t}$.

### 3.3.2 Semantic Knowledge Augmentation

#### Data Augmentation

To enable the model to detect unseen relations without labeled training instances, we first do data augmentation by translating a sentence from its original seen relation to a new unseen relation using an analogy. In the word level, we adopt 3CosMul [100], where we use

the top 10 similar words to return, to get the candidates of new words $w_u$:

$$w_u = \underset{x \in V}{argmax} \frac{cos(x, r_u) \cdot cos(x, w_s)}{cos(x, r_s) + \epsilon} \tag{3.1}$$

where $V$ is the vocabulary set, $cos(\cdot)$ is the cosine similarity, $r_u$ is the unseen relation, $r_s$ is the seen relation, $w_s$ is the word in seen class and $\epsilon$ is a small number to prevent division by zero.

**Algorithm 3** Sentence Generation for Unseen Relations

**Input** : sentence $x_i = [w_1^i, \cdots, w_n^i]$, two name entities $e_{i_h}$ and $e_{i_t}$, original relation label sets $R_s$, target unseen relation label $r_u$
**Output** : sentence $x_i^u$ with relation $r_u$

```
for r_s ∈ R_s do
    if S(r_u) == S(r_s) and S(e_u) == S(e_s) then
        for w ∈ x_i do
            if is_valid_pos(w) then
                w_u = 3CosMul(w, r_u, r_s)
                x_i^u.append(w_u)
            else
                x_i^u.append(w)
    else
        Continue
return x_i^u
```

At the sentence level, we follow Algorithm 3 to translate a sentence of relation $r_s$ into a new sentence of relation $r_u$. To be more specific, we translate all nouns, verbs, adjectives, and adverbs in the seen sentence to a new sentence. We do the translation when the super-class of $r_s$ and the super-class of two corresponding name entities in $r_s$ are the same as the super-class of $r_u$ and the super-class of two related name entities in $r_u$. If the number of $r_s$ that conforms to the rules is larger than one, we take all the translated sentences and randomly select the same number as other seen relations to make a balanced training set.

### Prompts from Knowledge Graph

For relation extraction, the core issue is to extract the possible triplets from all aspects and granularities. For zero-shot tasks, we design prompts used as training instances to help train the model because there is no real training data available. We construct prompts based on the external knowledge graph ConceptNet [194], a knowledge graph that connects words and phrases of natural language with labeled edges, for zero-shot relation extraction. Nodes in ConceptNet are entities, and edges connecting two nodes are semantic relations between the entities. Because of the relation extraction task, we wrap the input sequence with a

*template*, which is a piece of natural language text. To be more specific, we build prompts as '$S(e_{i_h})$ is [MASK] of $S(e_{i_t})$'. We consider different locations of prompts such as before and after the input sentence. There is a similar performance, so we put the prompts after each input sentence. The [MASK] here is a virtual label word $r_v$ representing the relation between $S(e_{i_h})$ and $S(e_{i_t})$. Unlike using real words, we build the virtual label word that can primarily represent the relation in each sentence. Instead of building a virtual label word by simply using the mean vector of the top_k high-frequency words [135], we build our virtual label word based on a knowledge graph using the following strategy.

**Algorithm 4** Virtual Label Generation

**Input :** word $w_i$, relation $r_c$, threshold $\tau_s$, number of hop $K$, Knowledge Graph $G$, number of virtual label $n$

**Output :** virtual label $r_v$

**for** $w_i \in V$ **do**
  **if** $\frac{w_i \cdot r_c}{|w_i| \times |r_c|} \geq \tau_s$ **then**
    $v_1 = 0$, $v_2$, $v_3$, $v_{ave} = []$
    **if** $w_i \in G$ **then**
      $v_1 = 1$
    **else**
      $v_1 = 0$
    **for** $k \in K$ **do**
      hops = find_neighbors($w_i$) $\in G$
      **if** *hops* **then**
        $v_2$.append(any(hops))
        $v_3$.append(sum(hops))
        $v_{ave}$.append(mean(hops))
      **else**
        $v_2$,$v_3$,$v_{ave}$.append(0)
      $\alpha_{w_i} = \frac{\sum v}{Dim(v)}$
  **else**
    Continue
  $\gamma_v = \frac{\alpha_{w_i} \cdot E(w_i) + \cdots + \alpha_{w_n} \cdot E(w_n)}{\sum \alpha}$
return $r_v$

We firstly represent a relation $r$ as five sets of nodes in ConceptNet by processing the class label $r_c$, class hierarchy $S(r_c)$, class description $D(r_c)$ and hierarchy of two name entities $S(e_{i_h})$ and $S(e_{i_t})$. We consider whether a word $w_i$ is related to the members of the five sets above within $K$ hops or not. The value of $K$ is determined through the grid search on the validation set. For each of the five sets above, we consider $v_1$ (whether $w_i$ is a node in $G$ in that set), $v_2$ (whether $w_i$'s neighbor is a node in $G$), $v_3$ (number of neighbors of $w_i$ in $G$). The above values associated with each set demonstrate the semantic distance of $w_i$ and the corresponding set. The detailed construction of virtual label $r_v$ is shown in Algorithm 4.

### 3.3.3 Model Architecture and Training

#### Instance Encoder

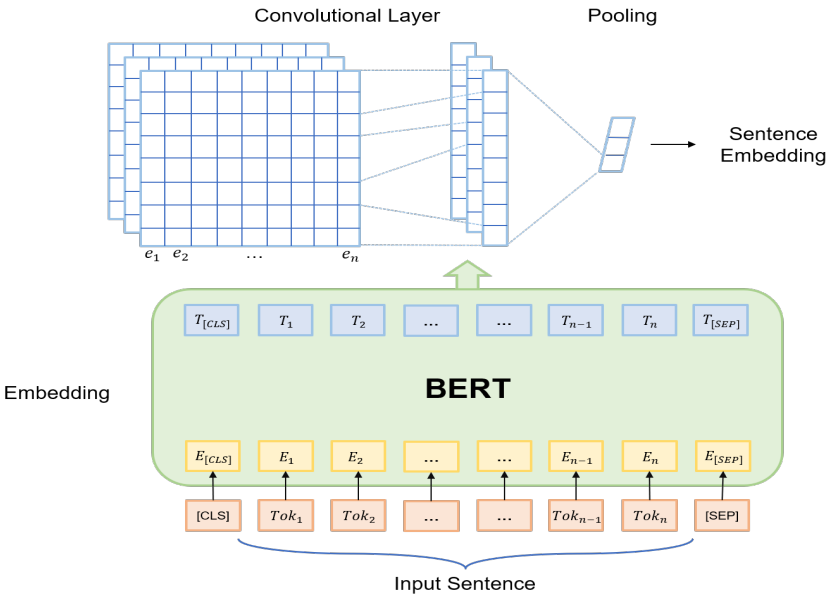


Figure 3.3: BERT-CNN Instance Encoder.

Figure 3.3 shows the architecture of the encoder used in this chapter. We first tokenize and lemmatize all words in a sentence. Two special tokens [CLS] and [SEP] are appended to the first and last positions, respectively. Then BERT [35] is used to generate the contextual representation for each token $w_i$. Because relation is not only related to the original name entities in augmented sentences generated by data augmentation, we have not used any position embeddings to show the positions of $e_{i_h}$ and $e_{i_t}$. Let $h_i$ represent the hidden state of the input sentence. We use $CNN(\cdot)$ ReLU and a max-pooling layer $max(\cdot)$, to derive the representation:

$$h_i = max(ReLU(CNN(x_i))) \tag{3.2}$$

where $x_i$ is the tokenized input sentence:

$$x_i = w_{i-\frac{n-1}{2}}, \cdots, w_{i+\frac{n-1}{2}} \tag{3.3}$$

We obtain the hidden state vectors of prompts $h_p$:

$$h_p = E(S(e_{i_h})) \oplus E(r_v) \oplus E(S(e_{i_t})) \tag{3.4}$$

where $E(\cdot)$ is the embedding function, $S(\cdot)$ is the super-class of the input word and $r_v$ is the virtual label embedding. The final representation for each instance is the concatenation of $h_i$ and $h_p$.

### Name Entity Extractor

ZS-SKA includes a name entity recognition encoder to extract name entities given the input $X_i$ with predicted relation $r_i$. A fine-tuned BERT [1] is implemented to recognize different types of entities. Each entity $e_i$ will be assigned a possibility score $score_i$ with entity types $T(e_i)$ (i.e. B-PER, I-LOC).

$$< e_i, T(e_i), score_i >= E_{NER}(X_i) \tag{3.5}$$

where $E_{NER}$ is the name entity extractor based on BERT. Based on the predicted relation sets $R_i$, super-classes of the possible name entities $S(e_{i_h})$ and $S(e_{i_t})$ can be accessed from the prompt templates. Then, we filter out the entities whose types are different from the super-classes in the prompt template for each relation:

$$E_i = \{e_i | T(e_i) = S(e_i), score_i >= \tau_e\} \tag{3.6}$$

where $E_i$ is the possible entity sets after the filter, $S(e_i)$ is the super-class of the target relation in the prompt template and $\tau_e$ is the threshold for the possibility of name entity types.

### Model Training

The objective of training ZS-SKA is to minimize the distance between each instance embedding $h_i \oplus h_p$ and the prototype $c_i$ embedding representing each learned relation (different colors in prototypes representation in Figure 3.2). Instead of using a softmax layer to classify seen relations and unseen relations, we adopt prototypical networks to compute a prototype for each relation after the BERT-CNN encoder. Each prototype is the average instance embeddings belonging to one relation:

$$c_i = \frac{1}{N} \sum_{i=1}^{N} f_\phi(h_i \oplus h_p) \tag{3.7}$$

where $c_i$ represents the prototype for each relation, $f_\phi$ is the BERT-CNN encoder, $h_i$ is the representation for each original or augmented sentence and $p_i$ is denotes the prompt embeddings introduced in Sec. 3.3.2. The probabilities of the relations in $R_s$ and $R_u$ for a query instance $x$ is calculated as:

$$p_\phi(y = r_i | x) = \frac{exp(-d(f_\phi(h_i \oplus h_p), c_i))}{\sum_{j=1}^{|R|} exp(-d(f_\phi(h_i \oplus h_p), c_j))} \tag{3.8}$$

where $d(.)$ is the Euclidean distance for two vectors. For multiple zero-shot RTE, we set a distance threshold $\tau_d$ to determine the number of possible unseen relations. During the

[1] https://huggingface.co/dslim/bert-base-NER

Table 3.1: The statistics of each dataset.

| | #instances | #relations | avg. len. |
|---|---|---|---|
| FewRel | 56,000 | 80 | 24.95 |
| Wiki-ZSL | 94,383 | 113 | 24.85 |
| NYT | 134,152 | 53 | 38.81 |

inference phase, ZS-SKA predicts the relation set $R_i$ by comparing the normalized distance $d(x_i)$ with the threshold $\tau_d$:

$$R_i = \{r_i | softmax(d(x_i)) < \tau_d, r_i \in R\} \tag{3.9}$$

where $d(.)$ is the Euclidean distance. The final distance threshold $\tau_d$ for the testing phase is chosen by the threshold value that has the best performance in the evaluation phase. For zero-shot RTE, the final relation triplets are the combination of $E_i$ from Equ. 3.6 and $R_i$ from Equ. 3.9. For zero-shot RC, only $R_i$ is provided for the result.

## 3.4 Experiments

We conduct several experiments with ablation studies on three public datasets: FewRel [72], Wiki-ZSL [16, 193] and NYT [173] to show that our proposed model outperforms other existing state-of-the-art models, and our proposed model is more robust compared with the other models in zero-shot learning tasks.

### 3.4.1 Evaluation Settings

#### Dataset

In our experiments, we evaluate our model over three widely used datasets: FewRel [72], Wiki-ZSL [16] and NYT [173]. FewRel and Wiki-ZSL are two balanced datasets and NYT is an unbalanced dataset. The statistics of FewRel, Wiki-ZSL, and NYT datasets are shown in Table 5.1. We provide a more detailed description below:

- ***FewRel [72].*** The FewRel dataset is a human-annotated balanced few-shot RC dataset consisting of 80 types of relations, each of which has 700 instances.
- ***Wiki-ZSL [16].*** The Wiki-ZSL dataset is a subset of Wiki-KB [193], which filters out both the 'none' relation and relations that appear fewer than 300 times.
- ***NYT [173].*** The NYT dataset was generated by aligning Freebase relations with the New York Times Corpos (NYT). There are 53 possible relations in total. It is an unbalanced noisy dataset because all the relations have a different number of sentences.

### Baselines and Evaluation Metrics

For RTE in ZSL, we compare our proposed model ZS-SKA with six SOTA zero-shot RTE models: **TableSequence** [203], **NoGen** [29], **RelationPrompt** [29], **ZETT** [96] with two sizes of T5 models (T5-small and T5-base), and **PCRED** [99]. For the zero-shot RC task, We compare our proposed model to eight existing RC models on all three public datasets to evaluate the model's ability to detect unseen relations. For clean FewRel and Wiki-ZSL datasets, we compare our model with **CNN** [244], **Bi-LSTM** [253], **Attentional Bi-LSTM** [263], **R-BERT** [223], **ESIM** [21], **CIM** [175], **ZS-BERT** [16], and **NoGen** [29]. The eight baselines above are reported by [16] and [29]. We also compare the robustness of our model with the most SOTA re-implemented **ZS-BERT**, **NoGen** and **RelationPrompt**. For noisy NYT dataset, we compare our model with the re-implemented **CDNN** [244], **REDN** [104] and **ZSLRC** [61]. The evaluation metric for a single RTE is Accuracy (Acc.) because each sentence only includes one gold triplet. The evaluation metrics for multiple RTE are Precision (Pre.), Recall (Rec.), and F1-score (F1), because there are at least two gold triplets in the testing set. For RC evaluation, we also use Precision, Recall, and F1-score, similar to those used for the above baselines.

### Parameter Settings

Table 3.2: Parameter Settings

| Parameter | Value |
|---|---|
| Word Embedding Dimension | 768 |
| Hidden Layer Dimension | 300 |
| Sentence Max Length | 128 |
| Convolutional Window Size | 3 |
| Batch Size | 4 |
| Initial Learning Rate $\alpha$ | 0.01 |
| Weight Decay | $10^{-5}$ |
| Number of Hops $K$ | 1 |
| Similarity Threshold $\tau_s$ | 0.6 |
| Distance Threshold $\tau_d$ | 0.05 |
| NER Threshold $\tau_e$ | 0.5 |
| Number of Virtual Label $n$ | 5 |

We follow the experiment settings as [29] and [16] to enable zero-shot RTE and zero-shot RC tasks. We randomly select $m$ unseen relations and remove all the instances related to these $m$ relations in the training set to ensure that these $m$ relations have not appeared in training data. $m$ is varied to examine how performance is affected. For the hyperparameter and configuration of ZS-SKA, we implement ZS-SKA with PyTorch and optimize it with

an SGD optimizer. The initial learning rate is selected via the grid search within the range of $\{1e-1, 1e-2, 1e-3, 1e-4\}$ for minimizing the loss, the cosine similarity threshold is selected from 0 to 1 with step size 0.1. The distance threshold for determining the number of triplets in a given sentence is set to 0.05, which is explored in the validation set. NER threshold is selected from 0.1 to 0.9 with a step size of 0.2. Table 3.2 shows other parameters. We follow the early stopping strategy when selecting the model for testing. The model is evaluated on the validation set every 50 epochs. The time for training is around 6 hours depending on the computing resources. GPU with 16G memory is required for training.

### 3.4.2 Results and Discussion

#### Zero-shot Relation Triplet Extraction

**Main Results** Table 3.3 shows the results of both single and multiple RTE on FewRel and Wiki_ZSL in ZSL. The results of our proposed model are reported by the average of five runs. For single RTE, we observe that ZS-SKA significantly outperforms other baselines when m=5 and m=10. From Table 3.3, ZS-SKA demonstrates better performance on Wiki_ZSL than on FewRel, as it shows a greater increase in accuracy compared to the strongest baseline. This indicates that our proposed model is more robust and effective on the dataset with more classes as Wiki_ZSL has 113 relations and FewRel has 80 relations in all. For multiple RTE, ZS-SKA consistently achieves the best F1 score in all settings, except when m=5 on the FewRel dataset. Compared to other baselines, ZS-SKA achieves a relatively high precision score, resulting in a better F1 score. Downstream applications for RTE, such as building knowledge graphs using the extracted triplets, require high-quality data. A high precision and relatively low recall performance may result in missing some gold labels (i.e. missing links for the knowledge graph). However, a high recall and low precision performance (i.e. NoGen achieves the best recall performance in all settings) means that the model returns many results, but most of its predicted labels are incorrect compared to the gold labels. This may introduce much noise (i.e. a noisy dataset) to downstream tasks. We conjecture that the high precision performance of ZS-SKA is due to setting a relatively smaller distance threshold in Sec. **??** for determining the number of relation triplets. In future work, a dynamic threshold could be added to adjust to different datasets.

#### Zero-shot Relation Classification

ZS-SKA also supports the zero-shot relation classification task by providing two name entities in the training set. Recognizing unseen relations is mainly supported by semantic knowledge augmentation in Sec. 3.3.2. Therefore, we carry out relation classification experiments on NYT, FewRel and Wiki_ZSL datasets to better evaluate zero-shot ability of ZS-SKA.

Table 3.3: Results for Zero-Shot Relation Triplet Extraction.

| #unseen relations | Model | FewRel Single Acc. | FewRel Multi Pre. | FewRel Multi Rec. | FewRel Multi F1 | Wiki_ZSL Single Acc. | Wiki_ZSL Multi Pre. | Wiki_ZSL Multi Rec. | Wiki_ZSL Multi F1 |
|---|---|---|---|---|---|---|---|---|---|
| m=5 | TabSeq [203] | 11.82 | 15.23 | 1.91 | 3.40 | 14.47 | 43.68 | 3.51 | 6.29 |
| | NoGen [29] | 11.49 | 9.45 | **36.74** | 14.57 | 9.05 | 15.58 | **43.23** | 22.26 |
| | RelPrompt [29] | 22.27 | 20.80 | 24.32 | 22.34 | 16.64 | 29.11 | 31.00 | 30.01 |
| | $\text{ZETT}_{\text{T5-small}}$ [96] | 26.34 | 31.12 | 30.01 | 30.53 | 20.24 | 31.62 | 32.41 | 31.74 |
| | $\text{ZETT}_{\text{T5-base}}$ [96] | 30.71 | 38.14 | 30.58 | 33.71 | 21.49 | 35.89 | 28.38 | 31.74 |
| | PCRED [99] | 22.67 | 43.91 | 34.97 | **38.93** | 18.40 | 38.14 | 36.84 | 37.48 |
| | **ZS-SKA (ours)** | **32.86** | **57.50** | 26.24 | 36.04 | **44.00** | **66.70** | 27.24 | **38.68** |
| m=10 | TabSeq [203] | 12.54 | 28.93 | 3.60 | 6.37 | 9.61 | 45.31 | 3.57 | 6.4 |
| | NoGen [29] | 12.40 | 6.40 | **41.70** | 11.02 | 7.10 | 9.63 | **45.01** | 15.70 |
| | RelPrompt [29] | 23.18 | 21.59 | 28.68 | 24.61 | 16.48 | 30.20 | 32.31 | 31.19 |
| | $\text{ZETT}_{\text{T5-small}}$ [96] | 23.07 | 25.52 | 29.61 | 27.28 | 14.37 | 19.86 | 27.71 | 22.83 |
| | $\text{ZETT}_{\text{T5-base}}$ [96] | 27.79 | 30.65 | 32.44 | 31.28 | 17.16 | 24.49 | 26.99 | 24.87 |
| | PCRED [99] | 24.91 | 30.89 | 29.90 | 30.39 | 22.30 | 27.09 | 39.09 | 32.00 |
| | **ZS-SKA (ours)** | **34.03** | **60.48** | 23.22 | **33.28** | **26.40** | **45.38** | 29.27 | **35.30** |
| m=15 | TabSeq [203] | 11.65 | 19.03 | 1.99 | 3.48 | 9.20 | **44.43** | 3.53 | 6.39 |
| | NoGen [29] | 10.93 | 4.61 | **36.39** | 8.15 | 6.61 | 7.25 | **44.68** | 12.34 |
| | RelPrompt [29] | 18.97 | 17.73 | 23.20 | 20.08 | 16.16 | 26.19 | 32.12 | 28.85 |
| | $\text{ZETT}_{\text{T5-small}}$ [96] | 21.08 | 16.20 | 23.22 | 18.90 | 10.74 | 14.96 | 19.31 | 16.79 |
| | $\text{ZETT}_{\text{T5-base}}$ [96] | **26.17** | 22.50 | 27.09 | 24.39 | 12.78 | 19.45 | 23.31 | 21.21 |
| | PCRED [99] | 25.14 | 27.00 | 23.55 | 25.16 | **21.64** | 25.37 | 33.80 | 28.98 |
| | **ZS-SKA (ours)** | 23.86 | **37.29** | 19.13 | **25.29** | 20.26 | 31.23 | 27.20 | **29.19** |

Table 3.4: RC results with different $m$ values on NYT.

| | m=15 Precision | m=15 Recall | m=15 F1 | m=30 Precision | m=30 Recall | m=30 F1 |
|---|---|---|---|---|---|---|
| CDNN | 27.94 | 44.10 | 33.72 | 10.17 | 25.62 | 14.23 |
| REDN | 66.52 | 65.47 | 66.98 | 57.19 | 56.80 | 56.99 |
| ZSLRC | 96.06 | 93.84 | 93.59 | 94.81 | **90.46** | 89.76 |
| **ZS-SKA** | **96.23** | **94.68** | **94.42** | **95.91** | 90.38 | **91.27** |

Table 3.5: RC results (m=15) on Wiki-ZSL/FewRel.

| | Pre. | Rec. | F1 |
|---|---|---|---|
| CNN | 14.58/14.17 | 17.68/20.26 | 15.92/16.67 |
| BiLSTM | 16.25/16.83 | 18.94/27.62 | 17.49/20.92 |
| $\text{BiLSTM}_{\text{att}}$ | 16.93/16.48 | 18.54/26.36 | 17.70/20.28 |
| R-BERT | 17.31/16.95 | 18.82/19.37 | 18.03/18.08 |
| ESIM | 27.31/29.15 | 29.62/31.59 | 28.42/30.32 |
| CIM | 29.17/31.83 | 30.58/33.06 | 29.86/32.43 |
| ZS-BERT | 34.12/35.54 | 34.38/38.19 | 34.25/36.82 |
| NoGen | **54.45/66.49** | 29.43/40.05 | 37.56/**49.38** |
| **ZS-SKA** | 41.78/45.03 | **40.50/51.86** | **39.30**/46.99 |

**Results on Unbalanced Dataset** The experiment results on unbalanced dataset NYT by varying $m$ unseen relations are shown in Table 3.4. To make fair comparisons, we use the same splitted NYT dataset and follow the same threshold schema provided by [61]. We remove the data augmentation module and only implement the prompts generated through the knowledge graph as similar side information in ZSLRC model. Apparently, the proposed ZS-SKA achieves a substantial gain in precision, recall, and F1-score over other baselines on the NYT dataset. When the number of unseen relations in the testing set becomes larger, the superiority of ZS-SKA gets more significant and robust. Such results indicate the effectiveness of leveraging prompts using virtual labels constructed from the knowledge graph instead of using keywords learned from the distribution of training data on the noisy dataset in [61].

**Results on Balanced Datasets** The evaluation results of zero-shot RC on Wiki-ZSL and FewRel are shown in Table 3.5. The results of all baselines are reported by [16, 29] and the result of ZS-SKA is reported by the average of five different random seeds. For a fair comparison, we compare our proposed model with baselines that do not require any training process for additional models such as the generator (GPT-2). Obviously, ZS-SKA significantly outperforms other existing models on both balanced datasets for recall value. Besides, ZS-SKA has the best F1 performance on Wiki-ZSL. The performance improvement indicates that semantic knowledge augmentation is competitively more beneficial for recognizing unseen relations than only incorporating text descriptions of relations.

### 3.4.3 Analysis

#### Ablation Study

Table 3.6: Ablation study (F1) over ZS-SKA on Wiki-ZSL with different percentages of unseen relations.

| | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|
| ZS-BERT | 58.31 | 19.59 | 17.63 | 11.79 | 9.52 |
| NoGen | 42.72 | 26.20 | 18.20 | 12.28 | 8.69 |
| RelPrompt | **67.91** | **50.02** | **36.51** | 22.13 | 12.94 |
| $\text{Ours}_{\text{Aug}}$ | 41.44 | 33.57 | 26.00 | 22.04 | 16.00 |
| $\text{Ours}_{\text{Prompts}}$ | 46.59 | 36.20 | 27.76 | 19.32 | 13.72 |
| $\text{Ours}_{\text{Top2 freq}}$ | 41.10 | 33.68 | 28.49 | 22.40 | 18.60 |
| $\text{Ours}_{\text{Top5 freq}}$ | 40.90 | 34.85 | 28.84 | 22.68 | 18.25 |
| $\text{Ours}_{\text{ActLabel}}$ | 42.06 | 34.89 | 28.85 | 22.93 | 18.43 |
| $\text{Ours}_{\text{OnlyBert}}$ | 37.35 | 31.80 | 25.80 | 20.75 | 16.51 |
| $\text{Ours}_{\text{All}}$ | 40.93 | 35.99 | 28.97 | **24.64** | **19.27** |

To evaluate the robustness and effectiveness of the zero-shot ability of ZS-SKA, we conduct an ablation study on Wiki-ZSL by removing different modules from ZS-SKA.The zero-shot setting is followed by the definition that partial relations are unseen in the testing set [216]. This setting is more competitive because all classes (including both seen and unseen relations) exist in the testing set. Different from the experiments of specific $m$ values, this is a 113-class classification experiment, including different percentages of unseen relations, which is more related to the real-world scenario. From Table 3.6, we observe that ZS-SKA is more robust when increasing the proportions of unseen relations. The performance drops drastically for ZS-BERT and NoGen when more unseen relations appear. RelationPrompt is more stable than ZS-BERT and NoGen. But the performance also drops a lot starting from 40% of unseen relations. Though instances generated by data augmentation for unseen relations may include noise, the models with data augmentation can be more robust when large percentages of unseen relations exist in the testing set. We also implement models with different ways to construct the prompt such as using top $k$ frequency words, and the actual label itself to evaluate the virtual label construction in ZS-SKA. Virtual label construction is more effective when 20% or more of unseen relations exist. It is because prompts constructed by virtual labels contain the semantic information of unseen relations, which shortens the distance between the query sentence of an unseen relation with the unseen relation prototype.

### Case Study

**Data Augmentation** Table 3.7 shows an example of the augmented data following the translating rule on the Wiki-ZSL dataset. The relation 'place_of_birth' is a seen class, and the other four relations are from unseen classes. We use data augmentation to generate augmented training instances for these unseen relations. We observe that if the super-class of both the relation and two name entities are the same, the generated sentences have a good

Table 3.7: Examples of sentence generation from seen relations by data augmentation. Words in red are name entities for each sentence. $S(\cdot)$ denotes the super-class of the relation or name entities.

| Relation $r$ | $S(r)$ | $S(e_1)$ | $S(e_2)$ | Sentence |
|---|---|---|---|---|
| place_of_birth | location | person | location | Jessica (born in Manchester) is a British track and field athlete who competes in the heptathlon. |
| place_of_death | location | person | location | Johnson (died in Liverpool) is a Military track and field athlete who competed in the decathlon. |
| residence | location | person | location | Mansion (resided in Villa) is a Colonial residence and peri alumnus who dominates in the decathlon. |
| country | location | location | location | Rich (retired in Arsenal) is a European track and field athlete who competes in the decathlon. |
| educated_at | act | person | org. | Jess (motivate in Liverpool) is a British aims and professional athlete who educated in the decathlon. |

quality with the name entities having unseen relations. If the super-class of the relation or two name entities of unseen relation is different from that of the seen relation, though the generated sentences contain the tone of the unseen relation (words in blue), the original two name entities do not have the target unseen relation. For example, the generated sentence of relation 'country' can be explained that Arsenal is from a European country, but such relation is lost between the two name entities 'Rich' and 'Arsenal'. Therefore, we follow the rule of using the relation and name entities from the same super-class with that of unseen relations to generate high-quality augmented instances for training in ZSL.

**Virtual Label Construction** Figure 3.4 shows an example of ranking the top ten components of the constructed virtual label before denoising and after denoising. The virtual labels shown in Figure 3.4 are generated by Algorithm 4. The red words are irrelevant to the relation 'religion_of'. After we refine the virtual label sets using the distance metric, these irrelevant words are filtered out in our virtual label sets, removing the noise in the knowledge graph.

**Name Entity Extractor** Figure 3.5 shows an example of how ZS-SKA uses the name entity extractor with super-classes information to extract relation triplets. ZS-SKA includes two steps for RTE. First, unseen relations are predicted by semantic knowledge augmentation. Based on the predicted relations, super-class information such as 'LOC' and 'PER' can be

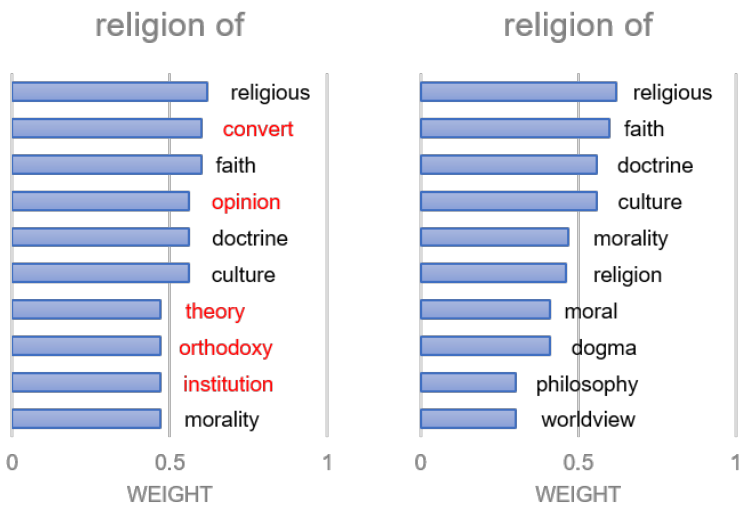


Figure 3.4: Denoising in virtual label construction.

accessed from the prompt template. Second, the NER extractor is implemented to extract the types of name entities. For example, the relation 'birthplace' happens between 'PER' and 'LOC' according to the template. The filter in the NER extractor selects 'Boyd' in 'PER' and 'Boston' in 'LOC'. Similarly, relation 'capital_of' only happens between 'LOC' and 'LOC', so the filter in the NER extractor selects 'Boston' and 'Massachusetts' in 'LOC'. Note that all locations in 'LOC' in Figure 3.5 are ranked based on the possibility score. Then, the predicted relations and entities extracted by the NER extractor construct the final relation triplets.

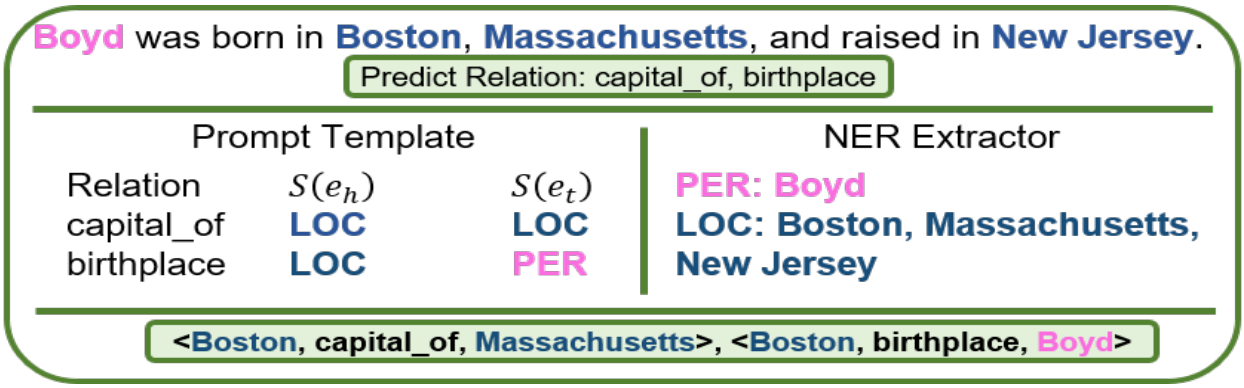


Figure 3.5: Example of using Name Entity Extractor to extract relation triplets.

### Hyperparameter Sensitivity

We examine how some primary hyperparameters, including threshold $\tau$ for denoising virtual label sets and the number of virtual labels $n$ in Algorithm 4 affect the performance of ZS-SKA. By fixing $m = 15$ and varying $\tau$ and $n$, the results in terms of F1 scores and Accuracy on NYT, FewRel and Wiki-ZSL datasets are exhibited in Figure 3.6. We find that parameters $\tau$ and $n$ affect the noisy dataset more than the clean and balanced dataset. We conjecture that because both $\tau$ and $n$ are used for removing noise and getting more related semantic information in prompts construction, the noise in prompts may impact more on noisy datasets because noisy datasets are more sensitive to the noise.

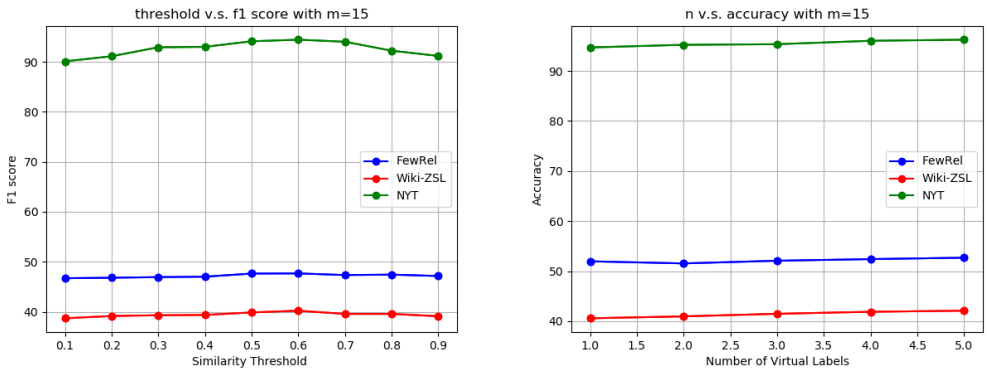


Figure 3.6: Effects on varying threshold $\tau$ and number of virtual labels $n$ on NYT, FewRel and Wiki-ZSL datasets.

If the threshold $\tau$ is between 0.5 and 0.6, it achieves the best performance on all three public datasets. This is reasonable that when $\tau$ is too low, most connected nodes in the knowledge graph are used to construct virtual label words. Thus, when building the prompts for each relation, it is more likely to bring the noise to the relation class. In contrast, when $\tau$ gets too high, some highly related nodes are filtered out to construct virtual labels. We would suggest setting $\tau$ between 0.5 to 0.6 to derive satisfying results across datasets. As for the number of words $n$ to construct virtual labels, we find that increasing the number of related words $n$ to construct virtual labels can achieve better performance. It is reasonable because, by including more nodes (words) from the knowledge graph to construct the virtual label representing the relation information, more semantic knowledge information is contained, leading to a shorter distance between the query sentence embedding with the prototype constructed from the prompts.

## 3.5 Limitations

Given the progress made to date with the work we propose in this chapter, we view the following current limitations as some opportunities to build on in future work. First, data augmentation is based on word-level transformation. With the development of generation models, more state-of-the-art data augmentation techniques can be implemented to generate data for zero-shot tasks in order to further improve performance. Second, the proposed prompt method depends on information from a fixed knowledge graph, which means it can not deal with the scenario if the unseen label is an out-of-vocabulary word. We have not considered this scenario because all classes from the three public datasets are well-known words or phrases. In future work, to get prompt information when the class word does not exist in the knowledge graph, we will consider directly using label descriptions or text generation models such as GPT-2 to generate label explanations.

## 3.6 Summary

In this chapter, we propose a ZS-SKA utilizing semantic knowledge augmentation to extract unseen relation triplets with no labeled data available for training to tackle with zero-shot RTE. The experiments show that with augmented instances, prompts generated through a knowledge graph, and a NER extractor with prompts, ZS-SKA outperforms other SOTA zero-shot RTE models. We have also conducted extensive experiments to study different aspects of ZS-SKA, from ablation study, and case study to hyperparameter sensitivity, and demonstrate the effectiveness and robustness of our proposed model. In future work, we plan to explore: (1) Different ways of instance generation and prompt designs for semantic augmented data. (2) Better approaches for constructing virtual labels in the prompt template. (3) More SOTA data augmentation techniques to generate data for zero-shot tasks to further improve the performance.

# Chapter 4

# Multi-Modal Few-Shot Relation Extraction with Hybrid Visual Evidence

The goal of few-shot relation extraction is to predict relations between name entities in a sentence when only a few labeled instances are available for training. Existing few-shot relation extraction methods focus on uni-modal information such as text only. This reduces performance when there is no clear contexts between the name entities described in text. We propose a multi-modal few-shot relation extraction model (MFS-HVE) that leverages both textual and visual semantic information to learn a multi-modal representation jointly. The MFS-HVE includes semantic feature extractors and multi-modal fusion components. The MFS-HVE semantic feature extractors are developed to extract both textual and visual features. The visual features include global image features and local object features within the image. The MFS-HVE multi-modal fusion unit integrates information from various modalities using image-guided attention, object-guided attention, and hybrid feature attention to fully capture the semantic interaction between visual regions of images and relevant texts. Extensive experiments conducted on two public datasets demonstrate that semantic visual information significantly improves performance of few-shot relation prediction.

## 4.1 Introduction

Relation extraction aims to predict the relation between two name entities in a sentence. To alleviate the reliance on high-quality annotated data, few-shot learning has drawn more attention, requiring only a few labeled instances for training to adapt to new tasks. Existing few-shot relation extraction methods can be roughly divided into two categories. One category involves methods only using plain text data, without any auxiliary information. For example, meta-learning models prototypical networks [50], siamese neural networks [243] are trained with only a few examples for each class to extract relations. The other category introduces external data sources such as relation information [130, 131], concepts of entities [236], side information [61], external datasets [56], and graphs [161], to compensate the limited information in the above methods, to enhance the performance in few-shot relation extraction.

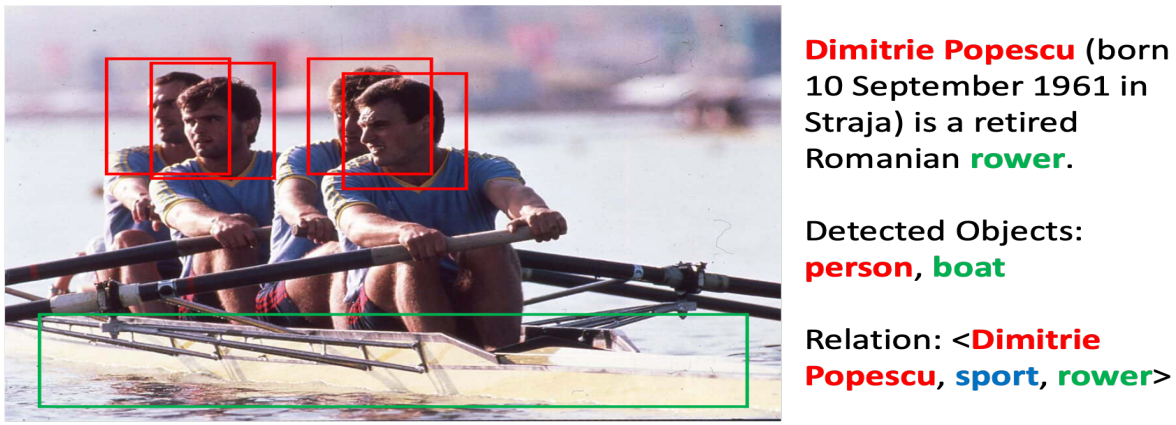


Figure 4.1: An example of multi-modal relation extraction based on visual information.

However, these methods mainly explore single-modality text-based data and may suffer a significant performance decline when texts lack contexts. For example, in Figure 4.1, given two name entities 'Dimitrie Popescu' and 'rower', it is difficult for text-based models to detect the relation 'sport' without other supplementary information because the word 'sport' or other similar words does not appear in the text. As a result, uni-modal models will incorrectly extract the relation 'winner' or 'candidate' of the two name entities according to the short given textual sentence. Even models using external information such as knowledge graphs or related words with similar meanings still can not correctly extract the relation due to the limited information in short given textual sentences.

Therefore, we question that *Can visual information be a good external source to supplement the missing contexts in textual sentences for few-shot relation extraction?* In the above case, we can easily classify the relation into 'sport' from the guidance of an image showing that a person is rowing a boat. Utilizing visual information to support contextual information for texts involves multi-modal learning. However, fusing information from different modalities is also a challenging task. First, simply concatenating textual and visual features without considering semantic information may even have a negative impact on the performance as shown in Sec. 4.4.4. For example, in Figure 4.1, the multiple people's faces in the background are noise for the image with the relation 'sport'. Second, existing multi-modal models (Sec. 4.2.2) mainly focus on fusing global visual features with text without considering the semantic information of visual objects in images. In Figure 4.1, visual objects such as 'person' and 'boat' contain essential information to the relation 'sport'.

To address these challenges, we propose a **M**ultimodal **F**ew-**S**hot model based on **H**ybrid **V**isual **E**vidence (MFS-HVE) for relation extraction. We first generate the representations through the textual feature extractor in Sec. 4.3.2 and the visual feature extractor in Sec. 4.3.2. We consider the visual representations from both the local perspective in low resolution (Sec. 4.3.3) and the global perspective in high resolution (Sec. 4.3.3). To be more specific, a local feature vector is the embedding of the objects detected from the image, and

a global feature vector is the embedding of the whole image. Because local features only focus on objects, global features can overcome the problem of sparsity with more information; however, they may probably contain noise (irrelevant information). We integrate both local features and global features to solve the problem of sparsity and noise.

Secondly, inspired by the cross-modal attention mechanism [241], we propose a multi-modal fusion unit including image-guided attention, object-guided attention, and hybrid feature attention to integrating semantic information from different modalities at both global and local levels. From the global perspective, image-guided attention based on the scaled dot-product attention [201] combines global feature vectors from the image with texts to capture the semantic interaction between visual regions of images and texts. From the local perspective, object-guided attention fuses objects detected from the image with relevant name entities from the textual sentences. Then the hybrid feature attention fuses all textual and visual information, including global image features and local object features. The hybrid feature attention generates a weight vector, multiplied by the multi-modal representations.

Finally, we concatenate text features, image-guided features, and object-guided features through a cross-modality encoder to generate the final multi-modal representations. Each relation representation is calculated based on the prototypical networks [191]. Next, based on the prototypical networks [191], we compute the mean value of all multi-modal support vectors as the prototype to represent each relation. Because of the hierarchical structure of the detected objects and name entities discussed in Sec. 4.3.3, hyperbolic distance is calculated between multi-modal query representations and prototypes to predict the relation. We conduct extensive experiments on two public datasets MNRE [258] and FewRel [72] to evaluate whether semantic visual information can supplement the missing contexts in textual sentences for few-shot relation extraction. FewRel is a uni-modal dataset containing only text, we crawl the image automatically by icrawler [1] for each instance to provide visual information, which can facilitate future research on multi-modal few-shot relation extraction. Details are introduced in Sec. 4.4.1. By comparing MFS-HVE with some state-of-the-art uni-modal few-shot relation extraction models and some multi-modal fusion methods with the same feature extractors, we show that, in general, models with multi-modal information perform better than the text-only models. However, our experimental results show that simply fusing all information directly without considering semantic contexts may have a negative impact. Our proposed model MFS-HVE, which integrates multi-modal semantic information at both global and local levels with three different attentions, outperforms other SOTA multi-modal fusion techniques introduced in Sec. 4.4.2. We also conduct ablation studies and parameter sensitivity studies to learn the impact of each attention and function. The contributions of this paper can be summarized as:

- We propose the first approach (MFS-HVE) for multi-modal few-shot relation extraction. Existing models for few-shot relation extraction only focus on a single data modality.

[1] https://icrawler.readthedocs.io/en/latest/

- MFS-HVE combines information from different modalities through image-guided attention, object-guided attention, and hybrid feature attention to integrating semantic visual information and textual information.
- We conduct extensive experiments on two public datasets. The experimental results show that introducing visual information can supplement the missing contexts in textual sentences for the few-shot relation extraction task.

## 4.2 Related Work

### 4.2.1 Few-shot Relation Extraction

Relation extraction predicts the relation of two name entities expressed in a sentence. Recent studies of few-shot relation extraction focused on metric-based representative methods. For example, the prototypical network learns a prototype for each relation via instance embeddings [7, 50, 87, 240]. Siamese neural network learns the metric of relational similarities between pairs of instances [52, 243]. Additional data sources are also used to help improve the performance in few-shot learning. Meta information such as relation information [37, 42, 112, 130, 131, 250, 261], concepts of entities [213, 236], additional auxiliary information [61], knowledge from cross domains [56], data augmentation [62, 159], and global graphs of all relations [161] are considered as prior information to establish connections between instance-based information and conceptual semantic-based information. However, the above studies only explore uni-modal text data. Different from these studies, we propose utilizing different data modalities, including both textual information and visual information, to supplement the missing semantics in texts.

### 4.2.2 Few-Shot Multi-Modal Fusion

Few-shot multi-modal fusion extracts relevant information from different modalities and integrates information collaboratively. MNRE is the first dataset developed for multimodal relation extraction [259]. Existing few-shot multi-modal fusion has been studied in the areas of visual question answering [91, 148, 199], image caption [2, 146], action recognition [152, 214], sentiment analysis [237], and so on. Studies have demonstrated that the performance of these tasks can be improved by fusing information from different modalities in few-shot learning [118]. Inspired by these works, we consider fusing visual information for few-shot relation extraction to provide the missing context in texts. The only work on few-shot relation extraction focuses on social relation extraction, in which relations describe connections only between people [202]. Besides, the dataset in [202] is not in English and includes a limited number of classes, and is therefore not sufficient to conduct 10-way-K-shot learning experiments. Considering the above limitations, we focus on few-shot general

relation extraction that is conducted on (1) a re-splitted MNRE dataset to satisfy few-shot learning, and (2) a subset of the FewRel dataset, where we collected corresponding images, to explore relation extraction in few-shot learning.

## 4.3 Methodology

In this section, we introduce the overview of MFS-HVE model. Figure 4.2 shows the architecture for few-shot relation extraction. It consists of two main modules: Semantic Feature Extractors and Multi-Modal Fusion. We describe these parts in detail below, starting with problem formulation.

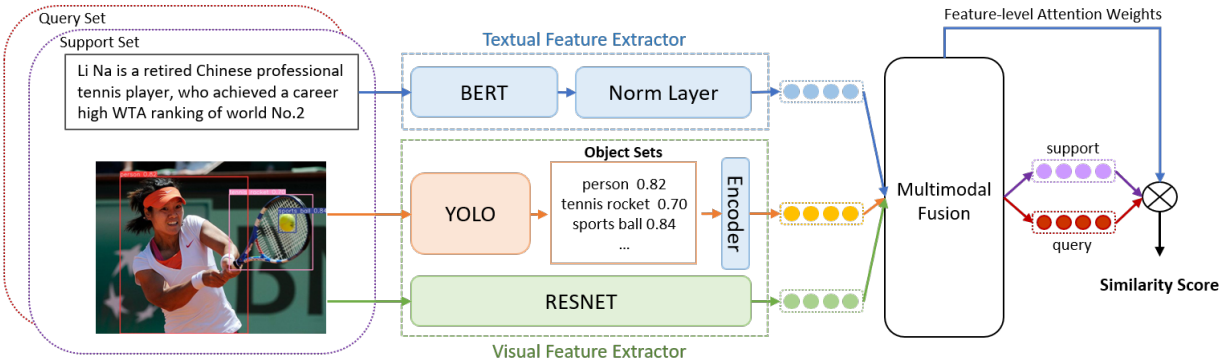


Figure 4.2: The overview of MFS-HVE. Details of multi-modal fusion is introduced in Sec. 4.3.3 and Figure 4.3

### 4.3.1 Problem Definition

We follow the N-way-K-shot definition and settings of few-shot learning from [50] to conduct our experiments. The N-way-K-shot setting means N classes with K examples of each. Typically K is no more than 10. There is no overlap between the classes in training data and testing data. For the multi-modal few-shot relation extraction task, we tend to classify the relation between two name entities based on text and image inputs. Let the input dataset represented by a set of tuples $(x_i, h_i, t_i, y_i, r_i)$, where $x_i$ is a sentence, $h_i$ is a head entity, $t_i$ is a tail entity, $y_i$ is the corresponding image and $r_i$ is the relation between $h_i$ and $t_i$. Our goal is to train a few-shot learning model $M$ to learn the representation function for the above tuples so that when randomly given support set with N relations and corresponding K tuples (NK tuples in total) as well as a query set with the same N relations and Q tuples, the model $M$ can predict the relations in the query set base on the given support set. $M$ is learned by minimizing the semantic distance between the input embedding from the support

set and the embedding from the query set. At test time, we use a different set of relations and evaluate performance on the query set, given the support set.

### 4.3.2 Semantic Feature Extractor

Each instance contains a text message and a corresponding image for relation extraction. The text is the input for the textual feature extractor, and the image is the input for the visual feature extractor.

#### Textual Feature Extractor

For the textual feature extractor, we use a pre-trained language model BERT [34] as the sentence encoder to generate the contextual representation. Two unique tokens [CLS] and [SEP] are appended to the first and last positions. The input text message is first tokenized into word pieces, and the positions of the name entities are marked by four special tokens [SEP] at the start and end of each entity mentioned in the relation statement of [7]. Then output representation of the textual feature extractor $r_t$ can be formulated as follows:

$$v_i = f_\phi(x_i, h, t) \tag{4.1}$$

$$r_t = tanh(W \cdot v_i + b) \tag{4.2}$$

where $v_i$ is the output of sentence encoder, $f_\phi$ is BERT encoder, $x_i$ is the input sentence, and $h$ and $t$ are head and tail entities, respectively. A fully-connected layer is added after BERT encoder, where $W \in \mathbb{R}^{256\times768}$ and $b \in \mathbb{R}^{256}$ are trainable.

#### Visual Feature Extractor

**Object Feature Representation** Object-level features are considered as the semantic information of the objects appearing in the image instead of the features of the whole image. For relation extraction tasks, a relation happens between the two name entities. Different from other multi-modal representation tasks, semantic information of the objects appearing in the images is of great importance. To extract objects from images, we utilize the pre-trained object detection model Yolo [10] to recognize the objects in the images. We consider the top $K$ frequent objects detected in the images to be the object labels because, in most cases, only the salient objects in the images are related to the name entities. Then, we transform the object labels into object embeddings to augment the semantic information of

the two name entities and address the problem of semantic disparity of different modalities. The representation of object-level features $r_o$ can be expressed as:

$$o = g_\phi(y_i) \tag{4.3}$$

$$r_o = f_\phi(o_0) \oplus \cdots \oplus f_\phi(o_k) \tag{4.4}$$

where $g_\phi$ denotes the object detection model, $y_i$ is the input image, $\{o_0, o_1, \cdots, o_k\} \in o$, indicating the objects detected in the image, $f_\phi$ is the object embedding encoder and $\oplus$ denotes concatenation.

**Image Feature Representation** The global image features are extracted from ResNet18 [75]. We use features from the last layer to produce the global vector. We then transform each feature vector into a new vector with the same dimension as the representation of the textual features using a single-layer perception. The representation of image-level features $r_i$ is:

$$v_i = h_\phi(y_i) \tag{4.5}$$

$$r_i = tanh(W \cdot v_i + b) \tag{4.6}$$

where $h_\phi$ denotes the image encoder, $y_i$ is the input image, $W \in \mathbb{R}^{256\times512}$ and $b \in \mathbb{R}^{256}$ are trainable weights and bias.

### 4.3.3 Multi-Modal Fusion

The architecture of our proposed multi-modal fusion is shown in Figure 4.3, including image-guided attention, object-guided attention, and feature-level attention.

#### Image-Guided Attention

A cross-modal attention layer can provide a more sophisticated fusion between different modalities [195]. Hence, we design a cross-attention layer module that combines the images and texts to capture the semantic interaction between visual regions of images and texts. As shown in Figure 4.3, the cross-modal attention layer is image-guided attention, which is calculated by combining Key-Value pairs from one modality with the Query from another modality. Specifically, the multi-modal representation is computed based on a modified version of the Scaled Dot-Product Attention (SA) [201]. The attended feature for images $\hat{f}_i = GA(q_i, k_t, v_t)$ is obtained by reconstructing $q_i$ using all samples in $v_t$ for their normalized cross-modal similarity to $q_i$. The image-guided attention unit is:

$$GA(q, k, v) = softmax((\frac{(W_Q q)(W_K k)^T}{\sqrt{d_k}})W_V v) \tag{4.7}$$

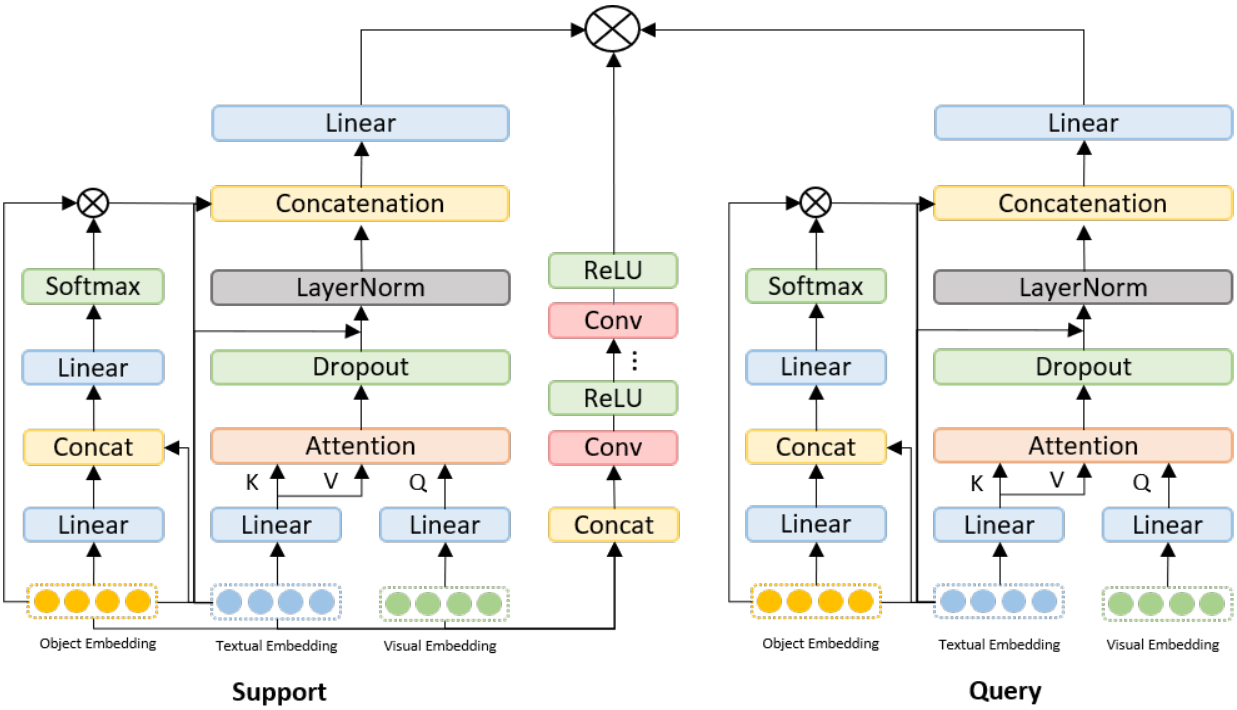


Figure 4.3: Detailed structure of multi-modal fusion.

where $W_Q$, $W_K$, $W_V$ are trainable query, key and value parameters and $d_k$ is the dimension of key vectors. Note that queries are from visual images, while keys and values are from text.

For each instance, the textual representation $r_t \in \mathbb{R}^{n \times d_t}$ is obtained through Equation 4.2 and image representation is obtained through Equation 4.6. We first input the textual representation $r_t$ and image representation $r_i$ into fully connected layers, respectively. Then, the image-guided attention unit models the pairwise relationship between the paired sample $< r_t, r_i >$, where $r_i$ guided the attention learning for $r_t$. The new image-guided feature vector related to $r_t$ based on the cross-modal attention can be expressed as:

$$\hat{r_i} = LayerNorm(r_t + GA(r_i, r_t, r_t)) \tag{4.8}$$

where LayerNorm is used to stabilize the training.

### Object-Guided Attention

Name entities in the textual sentence are always related to some objects detected from the input image. As shown in Figure 4.3, we propose an object-guided attention unit to fuse relevant words (name entities) and visual regions (objects). Given a textual feature $r_t$ obtained from Equation 4.2 and a local object feature $r_o$ obtained from Equation 4.4, we feed

these features into a single neural network layer followed by a softmax function to generate the attention distribution over the objects:

$$v_{r_t} = tanh(W_{r_t} r_t \oplus (W_{r_o} r_o + b_{r_o})) \tag{4.9}$$

$$\alpha_{r_t} = softmax(W_{a_t} v_{r_t} + b_{a_t}) \tag{4.10}$$

where $r_t \in \mathbb{R}^d$, $r_o \in \mathbb{R}^d$, $W_{r_o}$, $W_{r_t}$, $W_{a_t}$, $b_{r_t}$ and $b_{a_t}$ are all trainable weights and bias. $\oplus$ denotes concatenation. Based on the attention distribution $a_t$, the new object vector $\hat{r_o}$ related to $r_t$ is:

$$\hat{r_o} = \sum \alpha_{r_t} r_o \tag{4.11}$$

**Hybrid Feature Attention**

As shown in the middle of Figure 4.3, the hybrid feature attention fuses text information, global image-guided visual information, and local object-guided information, highlighting the important dimensions in the joint feature space to alleviate feature sparsity. For few-shot relation extraction, only a few instances in the support set are used for training so that the features extracted from the support set suffer from the problem of data sparsity. The feature-level attention generation block contains one concatenation layer, two or three 2D convolutional layers, and two or three activation functions, which can pay more attention to those more discriminative features when computing the space distance.

For space distance, studies show that hyperbolic spaces, where suitable curvatures match the characteristics of data, can lead to more generic embedding spaces [55, 128]. In the example shown in Figure **??**, the detected object 'person' is the hypernym of the name entity 'Magic Johnson' in the text. Thus, we adopt hyperbolic distance with feature-level attention in our networks to preserve such hierarchical structure:

$$d(s_1, s_2) = \alpha_i \cdot cosh^{-1}(1 + 2\frac{\|s_1 - s_2\|^2}{(1 - \|s_1\|^2)(1 - \|s_2\|^2)}) \tag{4.12}$$

where $\alpha_i$ is the score vector for relation $r_i$ calculated via the hybrid feature attention shown in Figure 4.3. By multiplying the hybrid feature attention weight by the support and query embeddings, we make the distance metrics better fit the given support sets and relations.

### 4.3.4 Model Training

The objective of training MFS-HVE is to minimize the distance between each instance embedding $L_{multi}$ and the relation embedding $P_{multi}(S)$. A cross-modality encoder concatenates the three vectors: sentence embedding $r_t$, object-guided textual embedding $\hat{r_o}$, and image-guided textual embedding $\hat{r_i}$, to yield the multi-modal representation. Then, a fully connected layer

is added to refine the multi-modal representation. The final multi-modal instance embedding $L_{multi}$ is:

$$L_{multi} = tanh(W_{multi} \cdot (r_t \oplus \hat{r_o} \oplus \hat{r_i}) + b_{multi}) \tag{4.13}$$

where $W_{multi}$ and $b_{multi}$ are trainable.

Given support set $S$ in the $N$ way $K$ shot setting, we compute a prototype for each of the $N$ relations $R$ in $S$ based on the multi-modal representations $L_{multi}$ of $K$ tuples. To be more specific, the prototype representation $P_{multi}(S)$ for $R$ is shown as:

$$P_{multi}(S) = \frac{1}{K} \sum_{i=1}^{K} L_{multi} \tag{4.14}$$

To predict the final relation among $N$ ways, hyperbolic distance $d$ as shown in Equation 4.12 is calculated between a query instance and each prototype $P_{multi}(S)$. Then, a softmax function is applied over the distance vector to generate a probability distribution on relations. More precisely, the probabilities of the relations for a query instance $q$ are computed as:

$$Pr(y = r_i|q) = \frac{exp(-d((L_{multi}), P_m(S)))}{\sum_{i=1}^{|R|} exp(-d((L_{multi}), P_i(S)))} \tag{4.15}$$

where $d(\cdot)$ is the hyperbolic distance.

## 4.4 Experiments

We conducted several experiments with ablation studies, case studies, and parameter sensitivity experiments on two public datasets: MNRE [258] and FewRel [72] to show that integrating semantic visual information with object-level and global feature-level attention mechanisms can help improve the performance, and our proposed multi-modal fusion method outperforms other existing fusion models.

### 4.4.1 Datasets

In our experiments, we evaluate our model [2] over two widely used datasets: MNRE [258], FewRel [72], and a subset of FewRel, which includes only clean images. FewRel is a balanced dataset, and MNRE is an unbalanced dataset. The statistics of MNRE and FewRel datasets are shown in Table 4.1. For the MNRE dataset, we randomly re-split the original supervised MNRE dataset to ensure that there is no overlap of relations between the training set and testing set. For FewRel and $\text{FewRel}_{\text{small}}$ datasets, we follow the same training and validation set. We describe each dataset and dataset construction in detail in the following:

[2]Code is available: https://github.com/gjiaying/MFS-HVE

Table 4.1: The statistics of each dataset.

| | #instances | #relations | avg. len. |
|---|---|---|---|
| MNRE | 15,484 | 23 | 16.67 |
| FewRel | 56,000 | 80 | 24.95 |
| FewRel_Small | 3,703 | 80 | 23.90 |

- ***MNRE [258].*** The MNRE dataset is a public human-annotated unbalanced multi-modal neural relation extraction dataset. It is originally built upon Twitter15 [134], Twitter17 [251] and crawling data from Twitter [3]. Each piece of data includes a sentence with two name entities and an image ID to correlate the text with the image. Because MNRE is a relation extraction dataset for supervised learning, there is an overlap of relations between the training and the testing dataset. For few-shot relation extraction, we randomly re-split the MNRE dataset to ensure no overlap of classes between the training and testing sets. There are 23 classes in total. After splitting the dataset, there are 13 classes for training and 10 classes for testing.

- ***FewRel [72].*** The FewRel dataset is a public human-annotated balanced few-shot RC dataset consisting of 80 types of relations (64 for training and 16 for validation, another 20 for testing but it is not public), each of which has 700 instances. Because we need to combine images with the original text, so we only run experiments on the public part (64 training + 16 validation). Because FewRel is a fully uni-modal dataset, we insert an image ID to each instance to make it into a multi-modal relation extraction dataset. The image for each instance is automatically crawled by a built-in web crawler [4] on wiki data from the Google search engine.

- ***FewRel$_{small}$.*** FewRel$_{\text{small}}$ is a subset of FewRel. Because FewRel doesn't have image information, we crawl the images for FewRel. We view these images as external information, similar to auxiliary information such as label description, knowledge graphs, entity description, etc. Because images crawled for FewRel is an automatic process, some of the images are not relevant to their corresponding texts. Noise exists in the newly constructed multi-modal FewRel dataset. Noisy images are removed to ensure that FewRel$_{\text{small}}$ is a small, clean, and high-quality multi-modal few-shot relation extraction dataset. Note that we did not do any labeling work. The labels remain the same in FewRel$_{\text{small}}$ as FewRel, and we only add more information (images) for the existing dataset.

In all, FewRel is a balanced dataset. Due to the data cleaning, FewRel$_{\text{small}}$ is an unbalanced dataset. MNRE is also an unbalanced dataset.

[3] https://archive.org/details/twitterstream
[4] https://github.com/hellock/icrawler

### 4.4.2 Baselines and Evaluation Metrics

We compare our model with six only text-based models: **Siamese** [98], **Proto** [191], **SNAIL** [145], **GNN** [180], **MLMAN** [240], **MTB** [7] and eight text-based models with external information: **REGRAB** [161], **ZSLRC** [61], **ConceptFERE** [236], **MapRE** [38], **HCPR** [71], **GM_GEN** [112], **FAEA** [42] and **SimpleFSRE** [131]. For multi-modal fusion baselines, we considering fusing the information from different modalities at different levels. The early fusion includes **Concatenation** [202], and **Circulant Fusion** [64]. The mid-level fusion includes **Deep Fusion** [212], **Dual Co-Att** [132], and **Proto$_{\text{multimodal}}$** [152]. We follow the same settings as [161] to run the experiments. The evaluation metric is the Accuracy (Acc.) of query instances.

### 4.4.3 Parameter Settings

Table 4.2: Parameter Settings

| Parameter | Value |
|---|---|
| Textual Information Dimension $d_t$ | 512 |
| Visual Information Dimension $d_v$ | 128 |
| Object Information Dimension $d_o$ | 256 |
| Batch Size | 1 |
| Initial Learning Rate $\alpha$ | 0.1 |
| Weight Decay | $10^{-5}$ |
| Dropout | 0.2 |
| Sentence Max Length | 128 |
| Objects Number | 2 |

For the hyperparameter and configuration of MFS-HVE, we implement MFS-HVE based on the PyTorch framework and optimize it with AdamW optimizer. We report the result based on a five-times run of the experiment. GPU of 16G memory is needed for the training process. The training time is around 5-6 hours depending on the computing resource. For the sentence encoder, we initialize the textual representation by pre-trained BERT [34] and set the dimension size at 768. Then we follow [7] to combine the token encodings of the entity mentioned in the sentence. For the image encoder, we initialize the visual representation by pre-trained ResNet18 [75] and set the dimension size at 512. For the object encoder, we employ 50-dimensional GloVe (6B tokens, 400K vocabulary) [155] for word embeddings of the objects detected from the image. Table 4.2 shows other parameters used in the experiment.

### 4.4.4 Results and Discussion

Table 4.3: Results of Accuracy Comparison Among Models (%) on MNRE and $\text{FewRel}_{\text{small}}$ Datasets.

| Modality | Model | MNRE | | $\text{FewRel}_{\text{small}}$ | | | |
|---|---|---|---|---|---|---|---|
| | | 5-Way 1-Shot | 10-Way 1-Shot | 5-Way 1-Shot | 5-Way 5-Shot | 10-Way 1-Shot | 10-Way 5-Shot |
| Only Text | GNN [180] | 29.08 | 22.53 | 46.38 | 70.45 | 28.74 | 62.07 |
| | Snail [145] | 30.90 | 19.43 | 40.16 | 60.07 | 21.19 | 47.56 |
| | Siamese [98] | 36.08 | 26.50 | 62.74 | 73.92 | 42.17 | 65.05 |
| | MLMAN [240] | 35.08 | 29.06 | 63.47 | 74.47 | 61.86 | 72.58 |
| | Proto_BERT [191] | 49.75 | 33.57 | 75.64 | 84.64 | 64.17 | 75.27 |
| | MTB [7] | 46.02 | 32.35 | 76.38 | 86.27 | 65.27 | 73.81 |
| Text+Others | ZSLRC [61] | 45.65 | 32.23 | 71.82 | 81.74 | 64.88 | 71.81 |
| | ConceptFERE [236] | - | - | 75.86 | 83.38 | 68.38 | 76.06 |
| | REGRAB [161] | - | - | 78.53 | 84.96 | 70.65 | 78.00 |
| | HCRP [71] | 31.10 | 10.45 | 78.04 | 84.68 | 69.54 | 77.91 |
| | MapRE [38] | 51.92 | 35.20 | 79.44 | 85.60 | 70.71 | 78.84 |
| | GM_GEN [112] | 52.58 | 35.82 | 60.04 | 73.74 | 42.22 | 59.23 |
| | FAEA [42] | 52.14 | 33.37 | 80.80 | 87.94 | 71.30 | 79.29 |
| | SimpleFSRE [131] | 50.32 | 35.05 | 80.84 | 87.46 | **71.67** | 80.14 |
| Text+Image | Concat [202] | 40.17 | 29.83 | 74.10 | 84.69 | 66.08 | 75.95 |
| | CirculantFusion [64] | 38.39 | 29.19 | 73.21 | 83.58 | 65.11 | 76.29 |
| | DeepFusion [212] | 48.27 | 33.28 | 78.38 | 86.76 | 66.36 | 76.08 |
| | $\text{Proto}_{\text{multimodal}}$ [152] | 50.84 | 34.10 | 77.18 | 86.28 | 68.19 | 78.29 |
| | Dual Co-Att [132] | 52.52 | 35.62 | 77.60 | 87.24 | 68.69 | 78.54 |
| | **MFS-HVE** | **54.88** | **36.62** | **81.32** | **89.65** | 69.52 | **80.55** |

### Main Results

The experiment results of few-shot learning on MNRE and $\text{FewRel}_{\text{small}}$ are shown in Table 4.3 with the average of five times run. Because some relations have less than ten instances in MNRE, it is impossible to run 5-shot experiments on MNRE, because we need five instances for the support set and the same number of instances for the query set. Thus, we only run 1-shot experiments on MNRE. FewRel is a public dataset with only textual information. We crawl the image relevant to each textual instance to construct a few-shot multi-modal dataset: $\text{FewRel}_{\text{small}}$, which is a subset of FewRel, including only clean images. Note that the baselines of multi-modal fusion works are implemented based on MTB [7] to have a fair comparison in few-shot relation extraction.

From Table 4.3, we observe that models integrating external information (labels, graphs, images, etc) perform much better than only text-based models. Models fusing semantic visual information can help improve the performance, but the performance highly depends on the fusion methods. Simply concatenating the visual information or fusing information at a coarse-grained level without considering semantic meanings such as circulant fusion may neg-

Table 4.4: Ablation study over MFS-HVE components (%) on MNRE and $\text{FewRel}_{\text{small}}$ datasets.

| Model Component | MNRE | | $\text{FewRel}_{\text{small}}$ | | | |
|---|---|---|---|---|---|---|
| | 5-Way 1-Shot | 10-Way 1-Shot | 5-Way 1-Shot | 5-Way 5-Shot | 10-Way 1-Shot | 10-Way 5-Shot |
| Only Text | 49.39 | 31.95 | 76.66 | 85.82 | 63.54 | 76.73 |
| Image Attention | 50.43 | 32.40 | 78.37 | 86.75 | 66.28 | 77.18 |
| Object Attention | 50.57 | 33.63 | 78.85 | 86.24 | 66.96 | 77.96 |
| Image&Object Attention | 52.26 | 35.38 | 80.50 | 88.72 | 69.49 | 79.17 |
| **MFS-HVE** | **54.88** | **36.62** | **81.32** | **89.65** | **69.52** | **80.55** |

atively impact the performance. This is probably because these methods treat all visual and textual information with equal importance (weights). However, only partial visual images contain relevant semantic meanings to the text. Directly using all information in the image may bring noise to the textual data. We further explore the robustness in Sec. 4.4.4. After considering image-guided textual information, object-guided textual information, and joint learning of text, image, and objects, our proposed model MFS-HVE significantly outperforms all state-of-the-art models on MNRE. More details about the performance of different attention layers of MFS-HVE are discussed in the ablation study in Section 4.4.4.

In summary, based on the experiment results on MNRE, FewRel, and $\text{FewRel}_{\text{small}}$, we have the following findings:

1. We find that models with multi-modal information perform better than text-based models in general.

2. Multi-modal models based on high-quality visual information are more robust than text-based models when the dataset size becomes smaller.

3. The performance of multi-modal models highly depends on the fusion methods. Simple concatenation or circulant multiplication of information from different modalities may probably have a negative impact.

4. For the relation extraction task, the local object information from the image is also very important because they are related to name entities in textual sentences and help reduce the noise of global image features.

### Ablation Study

To illustrate the effectiveness of MFS-HVE and explore the role of each attention unit in MFS-HVE, we carry out the ablation study on the datasets only with clean and high-quality

Table 4.5: Results of performance decrease in Accuracy(%) from FewRel to $\text{FewRel}_{\text{small}}$.

| Model | 5-Way 1-Shot | 5-Way 5-Shot | 10-Way 1-Shot | 10-Way 5-Shot |
|---|---|---|---|---|
| GNN [180] | 12.32 | 10.90 | 12.18 | 6.53 |
| Snail [145] | 9.88 | 12.12 | 11.19 | 11.58 |
| Siamese [98] | 5.61 | 8.36 | 14.24 | 4.25 |
| MLMAN [240] | 3.30 | 1.97 | 2.70 | 2.25 |
| Proto_BERT [191] | 2.20 | 4.31 | 3.11 | 7.35 |
| MTB [7] | 3.14 | 1.00 | 3.54 | 3.66 |
| ZSLRC [61] | 4.01 | 6.10 | 2.34 | 5.83 |
| ConceptFERE [236] | 3.56 | 2.96 | 3.34 | 3.76 |
| REGRAB [161] | 4.32 | 4.88 | 3.44 | 4.07 |
| HCRP [71] | 4.36 | 3.00 | 2.76 | 4.24 |
| MapRE [38] | 6.29 | 7.24 | 8.47 | 8.80 |
| FAEA [42] | 7.97 | 6.78 | 4.55 | 5.59 |
| SimpleFSRE [131] | 5.45 | 7.45 | 5.79 | 7.54 |
| Concat [202] | 3.08 | 1.32 | 2.58 | 0.83 |
| DeepFusion [212] | 2.14 | 4.72 | 0.38 | **0.39** |
| CirculantFusion [64] | 3.99 | 2.60 | 5.75 | 2.24 |
| Dual Co-Att [132] | 2.60 | 1.58 | 3.67 | 2.02 |
| $\text{Proto}_{\text{multimodal}}$ [152] | 2.01 | 3.08 | 3.75 | 2.99 |
| **MFS-HVE** | **1.95** | **0.83** | **0.27** | 1.32 |

visual data (MNRE and $\text{FewRel}_{\text{small}}$) because the performance of fusion with different multi-modal information is unstable with noisy data.The ablation experiment results shown in Table 4.4 are reported by the mean value of five times the experimental results. We observe that utilizing multi-modal information performs better than uni-modal information (text). However, only using image-guided attention or object-guided attention can not achieve a great performance improvement. This is probably because considering the whole image from a global perspective may introduce noise to the text, resulting in a similar performance in few-shot settings compared with text-based models. In addition, if only object-guided textual attention is added to the model, the model still can not achieve a significant improvement. This is because not all images include the objects that are relevant to the name entities in the text. Thus, when the model jointly fuses image attention and object attention, there is a promising performance increase. The image attention overcomes the problem of sparsity, whereas the object attention reduces the noise brought by the whole image features. After adding hybrid feature attention to fuse all textual and visual information from both global and local perspectives, a significant performance gain is seen.

### Model Robustness

To further study the robustness of integrating visual information with textual information, we also conduct experiments on the model's performance comparison on FewRel and $\text{FewRel}_{\text{small}}$. To make fair comparisons, instead of directly reporting the performance of other state-of-the-art models, we re-implement other models with the same parameter settings as the models run on $\text{FewRel}_{\text{small}}$. Table 4.5 shows the results of performance decrease from dataset FewRel to $\text{FewRel}_{\text{small}}$ in few-shot settings. Because the FewRel dataset is more than ten times larger than $\text{FewRel}_{\text{small}}$, there are more training instances in FewRel. It is reasonable to expect a performance drop when the model is training on a smaller dataset. From Table4.5, we observe that the performance of text-based models drops significantly when the dataset tends to be smaller. This is because models usually can perform better when more data is available. In addition, we also find that models based on multi-modal information are more robust than text-based models. They have a smaller performance decrease than text-based models. Our proposed model MFS-HVE performs the best in the one-shot learning setting. We conjecture that the high-quality semantic visual information neutralizes the negative impact of little training data in $\text{FewRel}_{\text{small}}$, resulting in a more robust performance of multi-modal models.

### Case Study

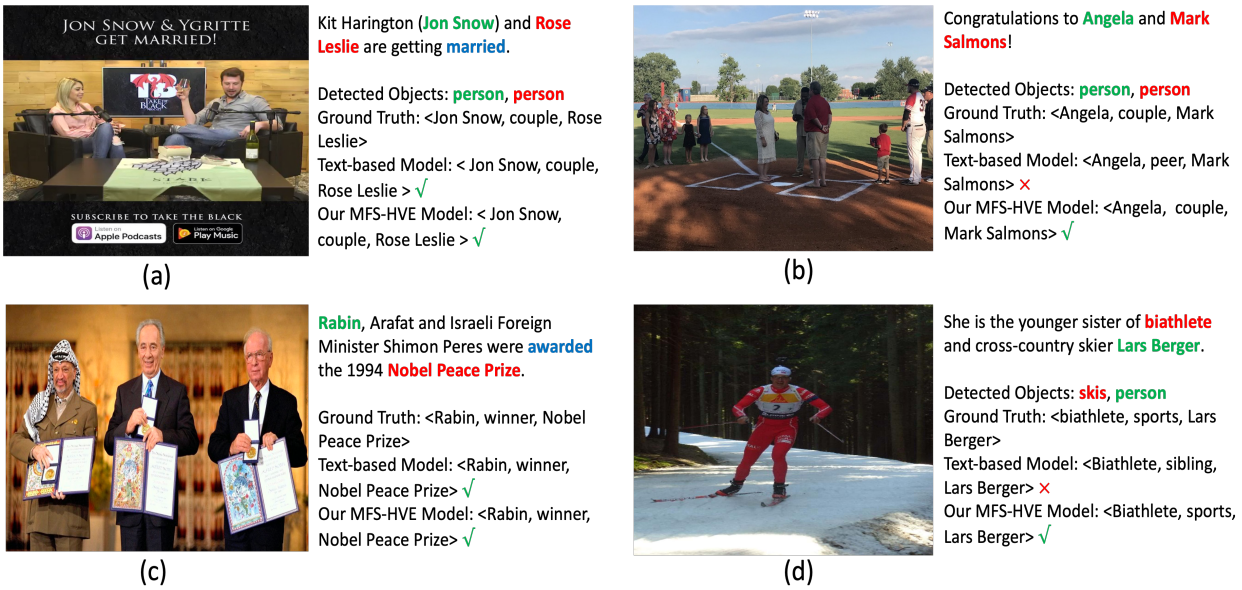


Figure 4.4: The examples of our proposed model MFS-HVE comparing to a text-based model on both the MNRE and FewRel datasets. We present the relation extraction results with the detected objects from the relevant image in the right column. The head entities are highlighted in green, whereas the tail entities are highlighted in red.

Figure 4.4 shows the case study comparing our MFS-HVE model with a text-based model MTB on both MNRE and FewRel datasets. To evaluate the advantage and effectiveness of semantic visual information, we compare our model with an unimodal model, which only depends on textual information. We present four examples of two relations. For each relation, we present two cases. One case is that both the text-based model and the multimodal model MFS-HVE predict the relation correctly. The other case is that the relation is incorrectly predicted by the text-based model but correctly predicted by MFS-HVE.

Based on these examples, we observe that the text-based model only performs well when rich information is in the text. For the examples shown on the left, the text-based model can only correctly predict the relation 'couple' when relevant words or phrases with similar meanings appear in the text, such as 'married' in the first sentence. Similarly, for the relation 'winner', the text-based model also performs well when the long textual sentence contains detailed information such as the word 'awarded'. These words relevant to the target relations provide enough semantic hints for the models with only text. However, not all cases have such long or detailed textual hints for the model. In the examples shown on the right, the textual sentences are short, without any words related to the target relation. In these cases, the text-based model can not predict the relation correctly. The text-based model predicts 'Angel' and 'Mark' are peers instead of 'couple', 'Roger Federer' is the 'participant of' the tennis tournament 'Wimbledon' instead of 'winner' of 'Wimbledon'. Nevertheless, with the guidance of informative visual evidence, more semantics are provided to the text. In the upper-right example, a wedding ceremony is shown in the image, and people objects are detected in the image. Based on this information, MFS-HVE correctly predicts the relation 'couple' instead of other relations in the MNRE dataset such as 'sibling', 'peer', 'parent', etc. Similarly, in the lower right example, MFS-HVE predicts the relation 'Roger Federer' is the 'winner' of 'Wimbledon' based on the visual information that a person is holding a tennis racket. In summary, integrating semantic visual information at both global and local levels provides more relevant information to supplement the missing contexts in textual sentences, resulting in a better and more robust performance for few-shot relation extraction.

### Parameter Sensitivity

Figure 4.5 shows the results of our proposed MFS-HVE model influenced by embedding a different number of objects detected from the image. By varying the object number from one to five, the results in terms of Accuracy on both MNRE and $\text{FewRel}_{\text{small}}$ are exhibited in Figure 4.5. We observe that the object number affects the performance of few-shot relation extraction. The model achieves the best performance when the object number is two. The performance drops when the object number increases. This is reasonable because relations always happen between two name entities. The two detected objects are usually relevant to the two corresponding name entities if the images are of high quality. Embedding only one object may lose critical information, whereas embedding lots more objects also introduces noise (irrelevant information) to the visual information.

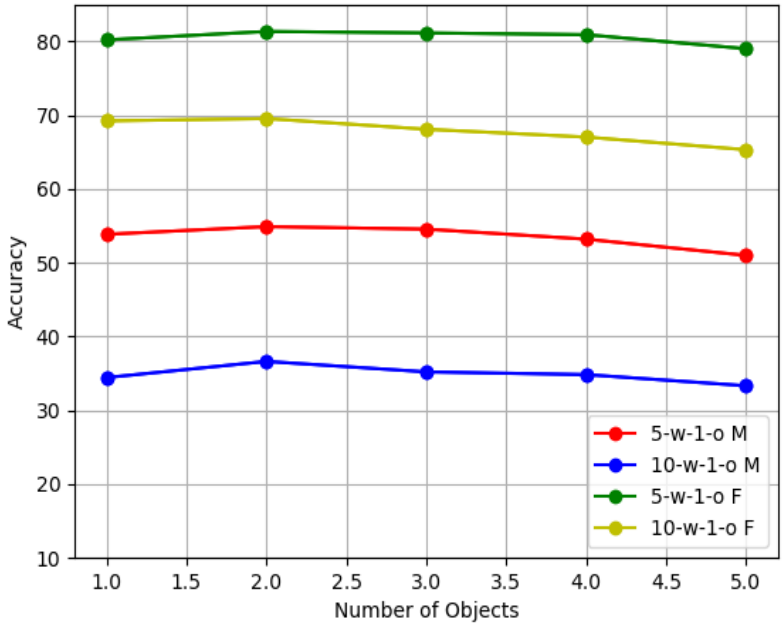


Figure 4.5: Effects on varying the number of embedded objects in one-shot settings on MNRE and $\text{FewRel}_{\text{small}}$ datasets.

## 4.5 Limitations

We view the following current limitations as some opportunities to build on in future work. First, MFS-HVE requires high-quality images for training. As shown in Table 4.3, MFS-HVE has a significant performance improvement compared with models using other text-based external information on MNRE. This is because MNRE is a public multi-modal dataset including clean and high-quality images. However, MFS-HVE shows a slight improvement or similar performance with models using other text-based external information on FewRel. The images crawled automatically contain much noise, which means some of the crawled images are irrelevant to the textual sentences. To further improve the performance on the FewRel dataset, human efforts or other crawling techniques are needed to get a large, clean, and high-quality image dataset.

Second, we compare MFS-HVE with five different fusion models introduced in Sec. 4.2.2. There are no existing multi-modal fusion models for the few-shot relation extraction task. We follow the five models' papers to implement the multi-modal fusion algorithms. To meet the requirement for few-shot learning, these fusion methods are built upon MTB [7]. More latest multi-modal fusion methods are needed for performance comparison. To further improve the performance, more SOTA visual encoders such as ViT [41] and large GPU memories are needed to conduct more experiments.

Finally, we want to clarify that our work focuses on few-shot relation extraction. We compare our model's performance with 14 SOTA open-code few-shot RE models and 5 different fusion models on two public English datasets. State-of-the-art multi-modal models in supervised learning for other tasks (i.e. NER, etc) or other languages besides English, are outside the scope of our paper because not all supervised models could be adapted/changed to few-shot settings as the training process is completely different.

## 4.6 Summary

In this chapter, we propose MFS-HVE, a multi-modal few-shot relation extraction approach leveraging semantic visual information to supplement the missing contexts in textual sentences. Our multi-modal fusion module consists of three attentions, visual-guided attention, object-guided attention, and hybrid feature attention that integrates information from different modalities at both global and local levels. Experimental results demonstrate that MFS-HVE leveraging attention-based multi-modal information outperforms other unimodal baselines and other state-of-the-art multi-modal fusion methods in few-shot relation extraction. We plan to explore the following directions in future work: (1) We will implement other powerful state-of-the-art image encoders such as ViT [41] to generate feature-level image embeddings. (2) Due to the performance improvement contributed by different attention layers in few-shot learning, we will explore utilizing the semantic visual information as an external source in zero-shot learning.

# Chapter 5

# Knowledge-Enhanced Multi-Label Few-Shot Product Attribute-Value Extraction

Existing attribute-value extraction (AVE) models require large quantities of labeled data for training. However, new products with new attribute-value pairs enter the market every day in real-world e-commerce. Thus, we formulate AVE in a multi-label few-shot learning scenario, aiming to extract unseen attribute value pairs based on a small number of training examples. We propose a Knowledge-Enhanced Attentive Framework (KEAF) based on prototypical networks, leveraging the generated label description and category information to learn more discriminative prototypes. In addition, KEAF integrates with hybrid attention to reduce noise and capture more informative semantics for each class by calculating both the label-relevant and query-related weights. To achieve multi-label inference, KEAF further learns a dynamic threshold by integrating the semantic information from both the support set and the query set. Extensive experiments with ablation studies conducted on two datasets demonstrate that our proposed model significantly outperforms other state-of-the-art models for information extraction in few-shot learning.

## 5.1 Introduction

Product attribute value pairs play important roles for e-Commerce because platforms make product recommendations for customers based on the key attribute-value pairs information and customers use attributes to compare products and make purchases. Existing studies on AVE based on neural networks view AVE as sequence labeling problems [89, 231], question-answering problems [187, 206] or multi-modal fusion problems [116, 266]. These supervised-learning models are well-trained to accurately classify attribute-value pairs when large quantities of labeled data are available for training. Even the most current open mining model needs a few attribute-value seeds and iterative training for weak supervision [254]. However, in the real world, new products with new attribute-value pairs enter the market every day in e-commerce platforms. It is difficult, time-consuming, and costly to manually label large quantities of new product profiles for training. Besides, with the appearance of new attribute-value pairs, the class distribution becomes long-tailed, where a subset of the

labels (head labels) have many samples, while the majority of the labels (tail labels) have only a few samples.

Given the above reasons, we formalize AVE as a multi-label few-shot learning (FSL) problem, aiming to extract structured product information from unstructured product profiles with limited training data. We take the common head labels data for training and the limited tail labels data for testing, and there is no overlap of classes between the training set and the testing set as shown in Figure 5.1. Recent methods on multi-label FSL have made great progress in computer vision [3, 189] and natural language processing [77, 123, 257]. Among these methods, prototypical network [192] has been proven to be powerful and has potential. However, different from the task for AVE in e-commerce, these models (1) explore only label tags as auxiliary information, (2) still have noise when learning prototypes, and (3) require further data or additional models to learn the threshold for label numbers prediction.

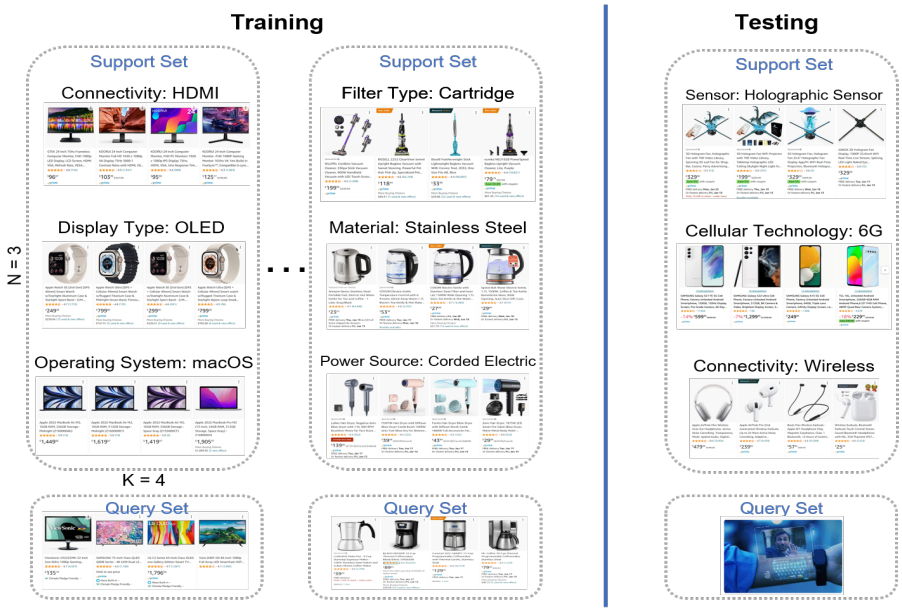


Figure 5.1: An example of multi-label few-shot product attribute-value extraction task.

To address the above challenges, we propose a Knowledge-Enhanced Attentive Framework (KEAF) for product AVE. The main contributions of KEAF consist of three parts. (1) To the best of our knowledge, we are the first to address AVE in the real-world scenario in e-commerce, where new products enter the market every day. We first formulate AVE as a multi-label few-shot task in Sec. 5.3.1 to tackle the problem of limited training data for long-tailed datasets. We sample and balance the real-world dataset in Sec. 5.3.2 to follow a multi-label few-shot setting. (2) By leveraging both the label description generated by a generator and the category information as the auxiliary information to obtain more discriminative prototypes, KEAF can not only avoid the issue that different attribute-value pairs share the identical prototype for 1-shot learning but also alleviate ambiguity by obtaining both label and category relevant information. The hybrid attention mechanism also helps reduce the noise and capture more informative semantics from the support set by calculating both the

label-relevant and query-related weights. (3) To achieve multi-label inference, a dynamic threshold is learned during the training stage by integrating the semantic information from support and query sets. The adaptive threshold does not require additional training data or is based on additional models. Extensive experimental results on two datasets show that our proposed model KEAF significantly outperforms other existing information extraction models for AVE.

## 5.2 Related Works

### 5.2.1 Attribute Value Extraction

Early works on attribute value extraction use a domain-specific dictionary and rule-based methods to identify attribute value pairs [59, 158, 186, 219]. With the development of neural networks, attribute value extraction has been modeled as a sequence labeling problem [89, 172, 231, 260]. To alleviate data sparse problems, several approaches introduce question-answering-based models for attribute value extraction [187, 206, 227]. Some methods fuse visual features to integrate more information for multi-modal attribute value extraction [116, 207, 266]. More recently, attribute value extraction has been viewed as an extreme multi-label classification task to reduce the size of labels through label masking [23]. However, these approaches require large quantities of data while training the model. When new products appear in real-world e-commerce platforms, it will be difficult for these approaches to accurately extract attribute-value pairs from product profiles with few labeled data.

### 5.2.2 Multi-Label Few-Shot Learning

Most works of FSL focus on single-label classification task [7]. However, for product attribute value extraction in e-commerce, one product may have multiple attribute-value pairs. Early works on multi-label FSL depend on a known structure of the label spaces [174] and label set operations [3].Then, prototypical networks [191] are revised for multi-label cases by learning a shared embedding space [238], grouping samples multiple times [189], and learning local features with different labels [233]. However, in one-shot learning, different classes may have the same prototype because one sample with multiple labels may contribute to the formation of several prototypes. To address this problem, recent works consider attention mechanisms [79] and label information to differentiate prototypes [77, 123, 257]. Different from these approaches, we leverage both label and category information for product attribute value extraction in e-commerce.

## 5.3 Methodology

### 5.3.1 Problem Definition

We consider the following multi-label few-shot classification setting. Given a set of training (base) classes $Y_{train}$ and testing (novel) classes $Y_{test}$, where $Y_{train} \cap Y_{test} = \varnothing$. The model is trained with numerous samples from $Y_{train}$, and it can quickly adapt to $Y_{test}$ with few labeled data. Each training episode involves a support set $S = \{(x_i, y_i)\}_{i=1}^{N_s}$ and a query set $Q = \{(x_i, y_i)\}_{i=1}^{N_q}$, where $S$ usually includes $K$ samples (K-shot) for each of $N$ labels (N-way).

Different from the N-way-K-shot setting of traditional single-label FSL [210], multi-label FSL allows each single instance can have multiple labels simultaneously. However, $N$ is the number of total classes, and each class has at least $K$ instances (at least one label will appear less than $K$ times with any instance removed) because we can not guarantee each label appears exactly $K$ times while each instance has multiple labels. Specifically, given a product data $x =< t, d, l, c >$, where $t$ denotes the product title consisting of $m$ tokens ($\{t_1, t_2, \ldots, t_m\}$), $d$ denotes the product description with $n$ tokens ($\{d_1, d_2, \ldots, d_n\}$), $l$ is the label description with $k$ tokens ($\{l_1, l_2, \ldots, l_k\}$), and $c$ denotes the product category. The input labels can be represented with a vector $y = \{y_1, y_2, \ldots, y_N\}$, where $y \in \{0, 1\}$, indicating the product has the label or not, and $N$ denotes the number of attribute-value pairs.

### 5.3.2 Multi-label Few-Shot Data Sampling

Data sampling for multi-label FSL includes three main steps: data splitting, data balancing, and data sampling. For data splitting, we reconstruct the dataset to guarantee that there is no overlap of classes between the training set $Y_{train}$ and testing set $Y_{test}$ for multi-label few-shot situations. We first set upper thresholds $t_u$ and lower thresholds $t_l$ based on the frequency of class labels for both the training set and the testing set. To solve the long-tailed problem by FSL, we set $Y_{train_{t_l}} >> Y_{test_{t_u}}$. Then, we filter the dataset by discarding the samples with the label count below $t_l$ or above $t_u$, updating the label dictionary, and discarding the samples with the label not in the label dictionary. To guarantee that the shot $K_S + K_Q >= 10$ for FSL, the filtering process is done iteratively until the number of classes $N$ is fixed.

Most data have only one label after data splitting, and a highly unbalanced dataset will have a bias when training and testing. For example, the model can achieve high performance when only predicting the label with the highest probability if most samples have only one label. To solve this problem, we balance the data by randomly dropping single-label data to reach a similar size with multiple-label data. Details of data are discussed in Sec. 5.4.1 and Table 5.1.

For data sampling, different from the single-label classification problem, it is difficult to ex-

actly follow the N-way-K-shot setting in multi-label FSL because each instance may contain multiple labels. To approximately conduct N-way-K-shot learning, we follow the setting [77]: (1) all labels appear at least $K$ times in each support set. (2) at least one label will appear less than $K$ times in the support set if any data pair is removed from it, to construct query and support sets for each episode. Details for multi-label few-shot data sampling are shown in Algorithm 5.

**Algorithm 5** Multi-label Few-shot Data Sampling

**Input :** Dataset $X$, label set $Y$, shot number $K_S$ for support and shot number $K_Q$ for query, upper threshold $t_u$ and lower threshold $t_l$

**Output :** Support set $S$, query set $Q$, query label set $Q_L$

Initialize $S = \{\ \}$, $Q = \{\ \}$, $Q_L = [\ ]$ and Dict $\{label : count\}$

```
while len(Y) is not fixed do
    if Count(Y_{X_{i,j}}) > t_u or Count(Y_{X_{i,j}}) <= t_l then
        X.remove(X_{i,j})
    Y.update(X)
    if get_class(X_{i,j}) ∉ Y then
        X.remove(X_{i,j})
data_balancing(X)
for i in Enumerate(Y) do
    indices = Random(Y_i, K_S + K_Q)
    count = 0
    for j in indices do
        if count < K_Q then
            Q.update(X_{i,j})
            Q_L.append(Y_{i,j})
        else
            if any Dict[Y_{i,j}] < K_S then
                S.update(X_{i,j})
                Update Dict
return S, Q, Q_L
```

### 5.3.3 Knowledge-Enhanced Attentive Framework

In this section, we introduce the overview of KEAF in Figure 5.2. It consists of four stages: contextual representation, label-enhanced prototypical network, hybrid attention, and dynamic threshold.

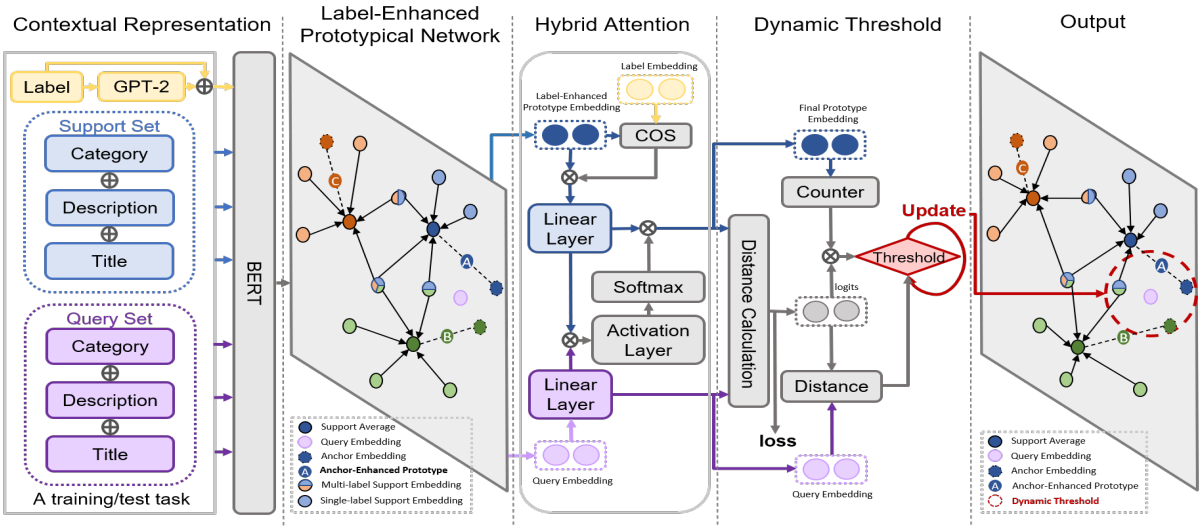


Figure 5.2: The overview of our proposed KEAF framework.

### Contextual Representations

Labels for AVE tasks are attribute-value pairs such as 'connectivity: wireless', which may lose contextual information due to the simple format. To achieve more information related to labels, we adopt GPT-2 [162] and Japanese GPT-2 [1] as the text generator to generate a detailed description for the attribute-value pairs. For the same example above, the generator can generate a more detailed explanation for the attribute-value pair as: 'connectivity is wireless communication between the user's device, which has an independent, physical signal to the user...'.

We adopt a pre-trained language model BERT [35] and Tohoku BERT [2] as the product input encoder to generate the contextual representation. We construct a string [CLS;$c$;SEP;$t$;SEP;$d$] by concatenating product category, title, and description as the input, where CLS and SEP are special tokens to represent a classifier token and a separator, respectively. The output representation for the product input $r_i$ and label input $l_i$ can be formulated as follows:

$$r_i = tanh(W \cdot f_{\varnothing}(c_i, t_i, d_i) + b) \tag{5.1}$$

$$l_i = tanh(W \cdot f_{\varnothing}(g_{\varnothing}(l_i)) + b) \tag{5.2}$$

where $f_{\varnothing}$ is BERT encoder, $g_{\varnothing}$ is GPT-2 generator, $c$ is category, $t$ is title, $d$ is description, $l$ is 'attribute is value' label information, $W$ and $b$ are trainable weights and bias.

[1] https://huggingface.co/rinna/japanese-gpt2-medium
[2] https://huggingface.co/cl-tohoku/bert-base-japanese

### Label-Enhanced Prototypical Network

As shown in Figure 5.2, we adopt prototypical networks [192] to get the original prototype representation of each attribute-value pair by averaging the embedding of support examples. However, different labels may share the same support instances in multi-label settings. In multi-label 1-shot cases, prototypes with different labels can have exactly the same embedding, which causes severe ambiguity for prototype representations. To emphasize the difference between prototypes and reduce such ambiguity, we leverage more detailed label descriptions generated by GPT-2 [162] to fully express the semantic information for attribute-value pairs and help learn more representative prototypes. Label information has shown a significant effect on learning more discriminative prototypes for multi-label FSL tasks [77, 123, 257]. Therefore, we combine the semantic label information with the average of support instances to compute a label-enhanced prototype $c_i$ with an interpolation factor $\eta$:

$$c_i = \eta \times E(y_i) + (1 - \eta) \times \frac{1}{K_i} \sum_{j=1}^{K_i} E(x_i^j) \tag{5.3}$$

where $E(\cdot)$ is the encoder in Sec. 5.3.3, $x_i^j \in \{x|(x, Y) \in S \wedge y_i \in Y\}$ is the support instance labeled with $y_i$, $y_i$ is the attribute-value description, and $K_i$ is the number of shot in support set $S$. The combination of label description embedding and support embedding helps the prototypes better separate from each other.

### Hybrid Attention

The aim of hybrid attention is to select more informative instances by retaining attribute-value relevant information while eliminating the negative effect triggered by the noise. As shown in the third stage in Figure 5.2, we first capture the similarity weight $\alpha_i$ in the label by calculating the semantic similarity between the label-enhanced prototype embedding $c_i$ from Equ. 5.3 and the attribute-value description embedding $l_i$ from Equ. 6.6:

$$\alpha_i = cos(c_i, l_i) \tag{5.4}$$

$$\hat{c_i} = \alpha_i \times c_i \tag{5.5}$$

where $cos(\cdot)$ denotes the cosine similarity and $\hat{c_i}$ captures the class-relevant information. To further capture other informative semantics from query-related instances and reduce the effect of noise, we apply the instance-level attention, where each instance representation has a different importance factor $\beta_i$:

$$\beta_i = \frac{exp(L(\hat{c_i}) \times L(E(x_i^q)))}{\sum_{i'=1}^{K} exp(L(\hat{c_{i'}}) \times L(E(x_{i'}^q)))} \tag{5.6}$$

$$\hat{r_i} = \beta_i \times \hat{c_i} \tag{5.7}$$

where $L(\cdot)$ is the linear layer, $E(\cdot)$ is encoder from Equ. 6.5, $x_i^q$ represents the query instance and $\hat{r}_i$ is the final prototype. Now, the final prototype $\hat{r}_i$ contains label-relevant semantic information and it can be closer to the instances with features more related to queries.

#### Dynamic Threshold

As shown in Figure 5.2, we train the threshold value $\tau$ during the training stage by integrating the semantic information from both the support set and the query set. The thresholding function $T(\cdot)$ is calculated by the production of query label counter $\varphi(x_i^q)$ with the relevance score between the final prototype embedding $\hat{r}_i$ in Equ. 5.7 and query instance embedding $r_i^q$ generated from Equ. 6.5. The number of query labels is estimated by averaging the number of support labels of support instance $x_i$:

$$\tau = T(\varphi(x_i^q), S) = \frac{1}{N \times K} \sum_{x_i \in X} \varphi(x_i) \odot d(\hat{r}_i, r_i^q) \tag{5.8}$$

where $S$ is the support set, $N$ and $K$ denotes N-way-K-shot, $\varphi(\cdot)$ represents the label counter, $\odot$ is element-wise production, and $d(\cdot)$ is the distance function. The threshold is dynamically updated for each training epoch. During the evaluation and testing phase, the framework predicts the query instance label set $Y_i^q$ by comparing the distance $d_i^q$ with the threshold $\tau$ calculated from Equ. 5.8:

$$Y_i^q = \{y_i^q | d_i^q < \tau, y_i^q \in Y\} \tag{5.9}$$

The final threshold for the testing phase is chosen by the threshold value that has the best performance in the evaluation phase. The model is trained by repeatedly sampling training episodes from $Y_{train}$ with support set $S$ and a query set $Q$. The model parameters are updated using the following binary cross entropy (BCE) loss:

$$\mathcal{L} = \sum_{I \in Q} \sum_{i=1}^{N} y_i^I \cdot log\sigma(q_i^I) + (1 - y_i^I) \cdot log(1 - \sigma(q_i^I)) \tag{5.10}$$

where $Q$ denotes the set of instances from the query set of the current training episode, $N$ is N-way, $\sigma(\cdot)$ is the sigmoid function, and $y_i^I$ represents the ground truth.

## 5.4 Experiments

### 5.4.1 Experimental Setup

#### Dataset

We evaluate our model over two datasets: a large e-commerce platform in Japan, and MAVE [234], a public product dataset for attribute value extraction. To simulate the few-

Table 5.1: Comparison of our dataset with existing multi-label few-shot datasets.

| Dataset | Train/Val/Test #Instance | Train/Val/Test #Label | Multi-label Percentage |
|---|---|---|---|
| MS-COCO [117] | 97,600/-/24,400 | 64/-/16 | 38.81% |
| FewAsp [80] | 40,960/10,240/12,800 | 64/16/20 | 63.5% |
| TourSG [218] | 19,351/1,600/4,800 | 68/17/17 | 18.13% |
| StanfordLU [43] | 3,517/2,512/2,009 | 14/10/8 | 16.57% |
| ML-FSIC [232] | - | 8/6/6 | - |
| iMaterialist [68] | >1,000,000 | 65/15/40 | - |
| **MAVE** [234] | **29,458/-/2,049** | **45/-/17** | **45.25%** |
| **Ours** | **477,166/-/6,421** | **23/-/14** | **43.38%** |

shot situation, we reconstruct the two datasets into FSL settings, where there is no overlap of classes between the training set and the testing set. The detailed statistics of the two newly reconstructed datasets compared with other multi-label FSL datasets are shown in Table 5.1. The distribution of label counts is shown in Figure 5.3.

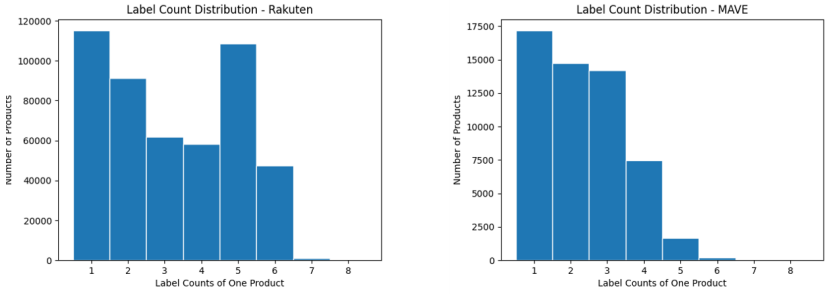


Figure 5.3: Label Count Distribution.

### Baselines and Evaluation Metrics

We compare KEAF with several state-of-the-art few-shot information extraction approaches with open code in two categories, only text-based models: **Siamese** [98], **Proto_BERT** [192], **MTB** [7], and models with label information: **HCPR** [70], **FAEA** [42] and **SimpleFSRE** [131]. They have achieved great performance for single-label FSL but failed to carry out multi-label few-shot prediction. To make them capable of multi-label prediction, and also to have a fair comparison, instead of directly using a softmax layer to output a probability vector, we add a dynamic threshold introduced in Sec. 5.3.3 for every baseline to predict the number of labels for each product. For evaluation metrics, we use both micro and macro precision, recall, and f1 score for multi-label few-shot prediction.

Table 5.2: Experimental results (%) of multi-label few-shot learning on an in-house E-Commerce dataset.

| Model | 1-shot | | | | | |
|---|---|---|---|---|---|---|
| | Macro | | | Micro | | |
| | P | R | F1 | P | R | F1 |
| Siamese [98] | 12.52 | 30.84 | 16.07 | 12.11 | 30.23 | 17.18 |
| MTB [7] | 13.00 | 36.86 | 17.70 | 12.77 | 36.76 | 18.87 |
| Proto [192] | 24.71 | 39.73 | 28.00 | **30.44** | 41.82 | 34.78 |
| HCRP [70] | 21.96 | 39.03 | 23.36 | 23.33 | 35.39 | 27.65 |
| FAEA [42] | 22.28 | 73.01 | 31.30 | 19.74 | 72.66 | 30.92 |
| SimpleFS [131] | 14.75 | 69.33 | 23.12 | 18.08 | 72.80 | 28.93 |
| KEAF w/o att | 24.79 | **73.57** | 34.43 | 24.54 | **75.35** | 36.92 |
| KEAF | **26.59** | 69.10 | **35.55** | 26.38 | 69.00 | **37.88** |

**Implementation Details**

For product and anchor embedding, we set the max length of product input as 512 and the max length of anchor description generated by GPT-2 [162] as 32. We initialize the contextual representation by pre-trained BERT [35] and set the dimension size at 768. We vary the label interpolation factor $\eta$ in $\{0.1, 0.5, 1.0\}$, and the optimal anchor weight is selected by grid search. For configurations, our model is implemented based on the PyTorch framework and optimized with AdamW optimizer. The learning rate is $10^{-5}$ with weight decay $10^{-6}$. The batch size is 1 and the dropout rate is 0.2. The experiments are conducted on Nvidia A100 GPU with 80G GPU memory.

### 5.4.2 Results and Discussions

**Main Results**

The experiment results of multi-label FSL on the two datasets are shown in Table 5.2 and Table 5.3, respectively. From these tables, we observe that: (1) Our proposed model significantly outperforms other baselines on both macro and micro F1 in 1-shot and 5-shot learning settings. These results reveal that our proposed model better learns the prototype representations and better captures the informative semantics. (2) On the in-house E-Commerce dataset, models using label semantics (HCRP, FAEA, SimpleFS, and KEAF) improve model performance more in a 1-shot setting than in a 5-shot setting. This is consistent with our expectations that adding label information helps reduce ambiguity, especially in 1-shot settings that only use support embeddings as prototypes. On MAVE, baseline models using label semantics even have worse performance than models not using label information. We conjecture that the original labels in MAVE are too simple for the models to learn the label

| Model | 5-shot | | | | | |
|---|---|---|---|---|---|---|
| | Macro | | | Micro | | |
| | P | R | F1 | P | R | F1 |
| Siamese [98] | 22.16 | 25.76 | 21.75 | 21.38 | 24.55 | 22.76 |
| MTB [7] | 10.16 | **98.89** | 18.06 | 10.12 | **98.51** | 18.35 |
| Proto [192] | 30.71 | 40.39 | 32.81 | **32.85** | 42.55 | 36.86 |
| HCRP [70] | 18.90 | 86.08 | 28.71 | 16.25 | 84.68 | 27.21 |
| FAEA [42] | 23.73 | 77.05 | 33.78 | 22.23 | 77.99 | 34.47 |
| SimpleFS [131] | 16.81 | 62.65 | 25.16 | 21.59 | 66.69 | 32.55 |
| KEAF w/o att | 26.39 | 73.50 | 36.69 | 26.01 | 75.08 | 38.48 |
| KEAF | **34.54** | 66.96 | **42.97** | 32.47 | 63.63 | **42.91** |

Table 5.3: Results of F1 score (%) on MAVE dataset.

| Model | 1-shot | | 5-shot | |
|---|---|---|---|---|
| | macro | micro | macro | micro |
| Siamese [98] | 15.83 | 18.55 | 28.29 | 28.15 |
| MTB [7] | 17.36 | 18.26 | 20.62 | 20.76 |
| Proto [192] | 30.09 | 36.79 | 33.56 | 39.46 |
| HCRP [70] | 18.55 | 19.31 | 16.11 | 16.58 |
| FAEA [42] | 16.38 | 16.87 | 16.63 | 17.11 |
| SimpleFS [131] | 26.63 | 30.43 | 17.44 | 23.05 |
| KEAF w/o att | 33.52 | 39.71 | **38.22** | 44.27 |
| KEAF | **34.47** | **44.46** | 36.40 | **44.33** |

Table 5.4: Ablation result over components in 1-shot learning setting on in-house E-Commerce and MAVE datasets.

| Model | In-house E-Commerce | | | | | |
|---|---|---|---|---|---|---|
| | Macro | | | Micro | | |
| | P | R | F1 | P | R | F1 |
| w/o anchor | 20.93 | 37.16 | 24.8 | 28.63 | 44.71 | 34.57 |
| w/o generator | 24.59 | 47.75 | 29.32 | 27.59 | 49.27 | 35.25 |
| w/o threshold | 24.64 | 44.66 | 29.37 | **28.68** | 48.26 | 35.57 |
| w/o category | 23.76 | 56.62 | 30.98 | 27.34 | 60.32 | 37.38 |
| w/o attention | 24.79 | **73.57** | 34.43 | 24.54 | **75.35** | 36.92 |
| KEAF (All) | **26.59** | 69.10 | **35.55** | 26.38 | 69.00 | **37.88** |

| Model | MAVE | | | | | |
|---|---|---|---|---|---|---|
| | Macro | | | Micro | | |
| | P | R | F1 | P | R | F1 |
| w/o anchor | 26.45 | 35.48 | 25.64 | 33.22 | 32.00 | 32.05 |
| w/o generator | 28.56 | 32.70 | 26.77 | 40.29 | 33.98 | 36.08 |
| w/o threshold | **32.99** | **61.88** | **38.22** | 33.45 | **66.77** | 44.27 |
| w/o category | 18.82 | 23.83 | 17.88 | 35.19 | 21.66 | 26.27 |
| w/o attention | 32.42 | 47.89 | 33.52 | 34.58 | 48.39 | 39.71 |
| KEAF (All) | 31.38 | 50.92 | 34.47 | 37.65 | 55.54 | **44.46** |

semantics. They even cause noise for the models. In our proposed model KEAF, we use the generator to generate a more detailed label description for better integrating the label information and reducing the noise, resulting in the best performance among all baselines. (3) MTB shows an extremely high recall score and other baselines also demonstrate good performances only on Recall. They all show a bad performance on Precision. For attribute-value extraction in e-commerce, a low precision score means lots of human efforts are needed to manually remove the non-relevant product attribute-value pairs. A possible reason for the large recall score is that the directly added dynamic threshold is not well-trained on these baselines for multi-label prediction. A very large threshold value is learned. These baseline models try to predict as many labels as possible, resulting in a very large recall value. In contrast, KEAF better learns the threshold and balances the precision and recall, resulting in the best performance on the F1 score.

### Ablation Study

To verify the effectiveness of each component in KEAF, we conduct a 1-shot ablation study on both two datasets by removing different components from KEAF in Table 5.4. We have the following observations based on Table 5.4: (1) Fusing anchor (label) information to

the prototype representations results in a significant performance improvement (i.e. 10.75% macro F1 increase on the in-house E-Commerce and 8.83% increase on MAVE) because the anchor information helps discriminate prototype embeddings and reduce ambiguity. (2) We observe that using the generator to generate a more detailed label description helps improve the performance more on MAVE (i.e. 11.7% increase of macro F1) than using the generator on the in-house E-Commerce (i.e. 6.2% increase of macro F1). We conjecture that this is because MAVE is an English dataset and English GPT-2 is better trained than Japanese GPT-2 using on the in-house E-Commerce dataset. The generated label description of MAVE is more accurate than the generated label description of the in-house E-Commerce dataset. (3) Knowledge of category information shows significant importance on MAVE (i.e. 18.19% micro F1 improvement). We think that this is because attribute-value pairs are from different categories and adding the category semantics can better separate the prototypes. (4) Using the attention mechanism can improve the performance by reducing the noise to some extent (i.e. less than 1% micro F1 increase on the in-house E-Commerce and 4.75% micro F1 increase on MAVE). More explorations on attention design are needed for future work.

## 5.5 Summary

In this chapter, we formulate an AVE task in the FSL scenario to solve the long-tailed data problem and limited training data for newly appeared products. We propose a knowledge-enhanced multi-label FSL method based on the prototypical network for product AVE. Specifically, we design the hybrid attention to alleviate noise and capture more informative semantics. Besides, we train a dynamic threshold to achieve multi-label inference. Extensive experimental results on two datasets demonstrate that our proposed method outperforms other state-of-the-art information extraction models significantly. Ablation study validates the effectiveness of knowledge enhancement and hybrid attention. In future work, we plan to explore (1) other knowledge such as taxonomy and images to get more semantic information, (2) contrastive learning to enlarge the inter-class difference for better prototype representation learning.

# Chapter 6

# Multi-Label Zero-Shot Product Attribute-Value Extraction

E-commerce platforms should provide detailed product descriptions (attribute values) for effective product search and recommendation. However, attribute value information is typically not available for new products. To predict unseen attribute values, large quantities of labeled training data are needed to train a traditional supervised learning model. Typically, it is difficult, time-consuming, and costly to manually label large quantities of new product profiles. We propose a novel method to efficiently and effectively extract unseen attribute values from new products in the absence of labeled data (zero-shot setting). In this chapter, we propose HyperPAVE, a multi-label zero-shot attribute value extraction model that leverages inductive inference in heterogeneous hypergraphs. In particular, our proposed technique constructs heterogeneous hypergraphs to capture complex higher-order relations to learn more accurate feature representations for graph nodes. Furthermore, our proposed HyperPAVE model uses an inductive link prediction mechanism to infer future connections between unseen nodes. This enables HyperPAVE to identify new attribute values without the need for labeled training data. We conduct extensive experiments with ablation studies on different categories of the MAVE dataset. The results demonstrate that our proposed HyperPAVE model significantly outperforms existing classification-based, generation-based large language models for attribute value extraction in the zero-shot setting.

## 6.1 Introduction

Product attribute value extraction (AVE) aims to extract attribute-value pairs (i.e. <color: red>) from e-Commerce product descriptions, which provides a better search and recommendation experience for customers. Existing studies on AVE mainly focus on supervised-learning models such as sequence labeling [89, 231], extractive question answering [187, 206] and multi-modal learning [60, 127, 204, 207] models. These supervised learning models are trained to only predict seen attribute value pairs. However, new products with unseen attribute-value pairs enter the market every day in real-world e-commerce platforms. It is time-consuming and costly to manually label large quantities of new products for training.

Some recent works focus on open mining models [228, 254] to directly extract attribute values

from product titles or descriptions. However, these approaches can not discover attribute values that are not explicitly mentioned in the text. In other words, these open mining models can not extract values that never appear in the product profile. To extract unseen attribute values, these open mining models use self-supervised learning, but they still need a high-quality seed attribute set bootstrapped from existing resources. Besides these open mining models, some generative large language models (LLM) are fine-tuned to autoregressively decode unseen attribute values from the input text. However, fine-tuning such LLM (i.e. T5 [163]) requires a lot of time and computing resources.

To address the above challenges, we propose HyperPAVE, a multi-label zero-shot attribute value extraction model that leverages inductive inference in heterogeneous hypergraphs to recognize unseen (new) attribute-value pairs (aspects) for which there is no available labeled training data. Motivated by the inductive graph learning, which shows the superiority of GNN to inductively adapt to infer unseen nodes [46, 226], we build inductive heterogeneous hypergraphs employing inductive link prediction mechanisms to infer missing or future connections (e.g., from new 'product' node to unseen 'aspect' node). The top part of Figure 6.1 shows an example comparison between supervised (Figure 6.1 a) and zero-shot (Figure 6.1 b) attribute value extraction. Existing works formulate relation propagation as a transductive link prediction task (Figure 6.1 a), where links can only be predicted between seen nodes (products and aspects) [18, 137]. To recognize unseen (new) aspects for new products, negative links are added in the original graph and the model is trained to predict whether an edge exists between two nodes based on the node features. HyperPAVE aims to learn the connections between both the nodes' features that are obtained from the fine-tuned LLM-based encoder and the complex graph structure. Motivated by the success of combining inductive GNNs and pre-trained BERT models [84], HyperPAVE is designed to enhance the inductive hypergraph-based model with fine-tuned BERT contextual embeddings for each node. Then, HyperPAVE is updated with zero-shot products and aspects with fine-tuned contextual embeddings, where message-passing is conducted directly on the updated graph, ensuring the inductive inference ability.

In addition, given the complexity of product data, it is important to design a model that can capture the heterogeneous, interconnected, and higher-order representation of both product data and user behavior data. Therefore, our proposed model HyperPAVE consists of various types of nodes including 'category', 'product', and 'aspect'. The product node records information including both product titles and descriptions. To fully express the semantic information for attribute-value pairs, the aspect nodes record detailed attribute-value descriptions generated by a generator. The proposed hypergraph representation uses higher-order relations to capture complex and interconnected user behavior information (e.g., 'also buy', 'also view') and product inventory information (e.g., 'product has aspects', 'category includes products'). The bottom part of Figure 6.1 shows an example comparison between graph-based (Figure 6.1 c) and hypergraph-based (Figure 6.1 d) attribute value extraction. To capture complex interconnected user behavior information, instead of using multiple graphs (one for each behavior e.g., "also buy" and "also view"), we construct hypergraphs

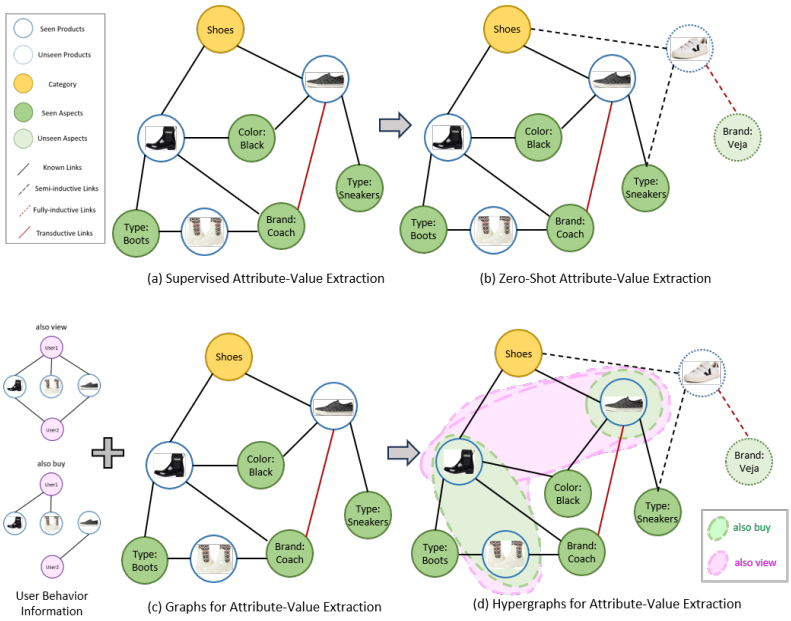


Figure 6.1: An example of zero-shot product attribute-value extraction by semi-inductive link predictions.

using hyperedges to represent user behavior information as higher-order relations. Compared to using several different graphs to capture complex relations, using a hypergraph (1) can include more (i.e. user behavior) information for the final node representation, (2) does not need to include user nodes in the graph, and (3) relations are not limited to binary connections. The contributions are summarized as:

- We propose a multi-label zero-shot model HyperPAVE to extract unseen attribute values for new products without labeled training data. HyperPAVE leverages an inductive link prediction mechanism combined with a fine-tuned BERT encoder to obtain unseen contextual node features.

- We build heterogeneous hypergraphs with higher-order relations to capture the complex and interconnected user behavior and structured product inventory information.

- Extensive experiments on the public dataset MAVE demonstrate that HyperPAVE significantly outperforms the classification model, generative LLMs, and graph-based models in zero-shot learning. Besides, HyperPAVE also shows the effectiveness and efficiency of training.

## 6.2 Related Works

### 6.2.1 Attribute Value Extraction

Attribute value extraction (AVE) aims at extracting attribute-value pairs (aspects) based on the product information. Early works use rule-based methods with domain-specific dictionaries to match target attribute value pairs [59, 158, 219]. With the development of neural networks, some studies view AVE as a sequence labeling problem [89, 172, 231, 260]. Then, question-answering-based models are built to treat attributes as questions and values as answers [187, 206, 227]. Multimodal fusion utilizing product images as visual features are learned to integrate visual semantics for products [30, 60, 116, 127, 129, 204, 207, 266]. Some studies formulate AVE as a multi-label classification task to extract multiple aspects for the products [23, 33, 63]. To handle unseen attribute values, open mining models [228, 254] extract aspects directly from the text with limited/weak supervision, and generation models [188] decode aspects as target sequences. However, all of these approaches (1) require large quantities of labeled data for training and (2) miss higher-order relations between products, such as 'also buy' or 'also view' products.

### 6.2.2 Zero-shot Learning

Zero-shot learning has been widely applied in the field of computer vision (CV) [157] and natural language processing (NLP) [17]. Existing works for zero-shot learning in information extraction can be roughly divided into three categories: (1) Embedding-based models, where representations of both seen and unseen classes are learned based on the auxiliary information such as class information [15, 179] and other external information [61, 122]. However, high-quality external knowledge is required for training the model, resulting in an increase in training time and resources. (2) Generative-based models, where augmented samples are generated for unseen classes by generation models (i.e. GAN [144], VAEs [97], and GPT-2 [162]) based on the samples of seen classes. Then, the zero-shot learning problem is converted into a conventional supervised learning problem [29, 62]. However, these models suffer from the noise of augmented samples and performance highly depends on generative models. (3) Graph-based models, where GNNs [181] are directly used to predict unseen classes by inductive link prediction [17]. Most studies view this problem as zero-shot knowledge graph completion [58] or zero-shot item recommendation [46]. Attentive GCN is used to transfer features from seen classes to unseen classes [65]. Ontologies or topologies are utilized to augment ZSL by capturing relationships between classes [27, 57]. Motivated by this, we build a product heterogeneous hypergraph to identify unseen aspects with inductive inference ability while capturing higher-order relations.

### 6.2.3 Heterogeneous Hypergraph

Hypergraphs are generalizations and extensions of ordinary graphs, where hyperedges can accommodate an arbitrary number of nodes to capture the higher-order relations [249]. To handle different types of nodes and edges, heterogeneous hypergraphs are learned by attention mechanisms [36, 95, 111, 124], wavelets [196], and variational auto-encoder [44, 125]. Though, all of these works are widely applied for social networks [113, 215], academic citations [217, 221, 248], biological networks [69, 141] or product recommendation in e-commerce [19, 28, 126, 229], heterogeneous hypergraphs are never applied to attribute value extraction in e-commerce. Different from the above hypergraphs that build hyperedges by close neighbors or meta-paths, we construct e-commerce related hyperedges by using user behavior and product inventory data to capture higher-order relations among categories, products, and aspects, to recognize unseen attribute values for new products.

## 6.3 Methodology

### 6.3.1 Problem Definition

In this section, we introduce the problem statement and some necessary definitions and notations for heterogeneous hypergraphs and multi-label zero-shot learning.

**Problem Statement.** Let $D = \{c_i, p_i, a_i\}$ denote a corpus of e-commerce product records, where $c_i$, $p_i$, $a_i$ represent sub-category, product and attribute value pair (aspect), respectively. We use $C$, $P$, and $A$ to denote the sets of sub-categories, products, and aspects. Hence, the task of attribute value extraction can be formulated as follows: Input: The product records $D$. Output: A model to estimate the probability that a new product $p$ in sub-category $c$ will have the unseen attribute value $a$. The goal of attribute value extraction is to learn a model $\mathcal{M}(p_i, a_j) \rightarrow \hat{y}\,[0, 1]$ to score the probability that a product $p_i$ has the attribute value $a_j$ based on $\mathcal{G}$, which includes all the relations from user behavior and product inventory information. Given several different graphs (i.e. user behavior graphs, product inventory graphs, etc.), we first build a heterogeneous hypergraph $\mathcal{G}$ to capture the higher-order and non-binary relations contained in $\mathcal{G}$. Then, we aim to learn the representations for nodes on a heterogeneous hypergraph $\mathcal{G}$ for an inductive link prediction task.

**DEFINITION 1 (Heterogeneous Hypergraph):** A heterogeneous hypergraph can be defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{T}_v, \mathcal{T}_e, W\}$, where $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$ is the node set, and $\mathcal{T}_v$ is the node type set. $\mathcal{E} = \{e_1, e_2, \cdots, e_M\}$ is the hyperedge set, and $\mathcal{T}_e$ is the hyperedge type set, where $|\mathcal{T}_v| + |\mathcal{T}_e| > 2$. $N$ and $M$ represent the maximum numbers of hyperedge nodes and edges. $W = diag(w_{e_1}, w_{e_2}, \cdots, w_{e_M})$ denotes the diagonal matrix representing the hyperedge

weight. We use incidence matrix $H \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{E}|}$ to represent relationships between nodes and hyperedges, with entries defined as:

$$H(v,e) = \begin{cases} 1 & \text{if } v \in e \\ 0 & \text{if } otherwise. \end{cases} \tag{6.1}$$

$D_v \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ and $D_e \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{E}|}$ are the diagonal matrices representing the degree matrix of nodes and hyperedges, where $D_v(i,i) = \sum_{e\in\mathcal{E}} W(e)H(i,e)$ and $D_e(i,i) = \sum_{v\in\mathcal{V}} H(v,i)$. The normalized hypergraph adjacency matrix $A \in \mathbb{R}^{\mathcal{V}\times\mathcal{V}}$, representing the connection relationship between nodes, is defined as:

$$A = D_v^{-1/2} H W D_e^{-1} H^T D_v^{-1/2} \tag{6.2}$$

**DEFINITION 2 (Zero-Shot Learning in Graph):** For multi-label zero-shot attribute-value (aspect) prediction, let $A^s = \{a_1^s, \cdots, a_m^s\}$ and $A^u = \{a_1^u, \cdots, a_m^u\}$ denote the node sets of seen and unseen aspects, where $A^s \cap A^u = \emptyset$. Only $A^s$ is included in the training graph $\mathcal{G}_{tr}$ and only $A^u$ is included in the testing graph $\mathcal{G}_t$. Product $p_i$ with any $a_i^u$ will be removed from $\mathcal{G}_{tr}$ to $\mathcal{G}_t$, to ensure all unseen aspect nodes are not in the training graph $\mathcal{G}_{tr}$. Details for multi-label zero-shot sampling are introduced in Algorithm 6.

**Algorithm 6** Multi-label Zero-shot Data Sampling

**Input :** Graph $\mathcal{G}$ with categories nodes $C$, product nodes $P$ and aspect nodes $A$, unseen aspect number $N$

**Output :** Train graph $\mathcal{G}_{tr}$, val graph $\mathcal{G}_v$, test graph $\mathcal{G}_t$

Initialize $\mathcal{G}_{tr}$, $\mathcal{G}_v$, $\mathcal{G}_t$

**for** *i in Random*$(N)$ **do**
- $P_i = get\_node(\mathcal{G}, A_i)$
- $link_{pos} = get\_edge(\mathcal{G}, P_i, A_i)$
- $link_{neg} = Sampling(get\_complement(link_{pos}))$
- $\mathcal{G}$.remove$(A_i, P_i, link_{pos})$
- **if** *i//2=0* **then**
  - $\mathcal{G}_v$.update$(A_i, P_i, link_{pos}, link_{neg})$
- **else**
  - $\mathcal{G}_t$.update$(A_i, P_i, link_{pos}, link_{neg})$
- $\mathcal{G}_{tr} = \mathcal{G}.add\_negatives()$

return $\mathcal{G}_{tr}$, $\mathcal{G}_v$, $\mathcal{G}_t$

### 6.3.2 Multi-Label Zero-Shot Data Sampling

Multi-label zero-shot data sampling includes (1) data splitting to ensure that there is **no overlap** of aspect and product nodes in training and validation/testing sets, and (2) negative

sampling to balance the dataset. For data splitting, we first randomly generate $N$ aspect nodes $A_N$ as unseen attribute values. Then, we remove both the nodes $A_N$ and their corresponding products $P_M$ as unseen products, and all edges on $A_N$ and $P_M$ from the original graph $\mathcal{G}$, where $N \neq M$. This step ensures that the zero-shot products and attribute values are never shown in the training graph. We update the validation and testing graphs with the zero-shot nodes and links separately so that there's no overlap of zero-shot nodes and links between the validation and testing sets. To balance the dataset, we do negative sampling and add negative links for all training, validation, and testing graphs. Details for multi-label zero-shot data sampling are shown in Algorithm 6.

### 6.3.3 Overall Framework

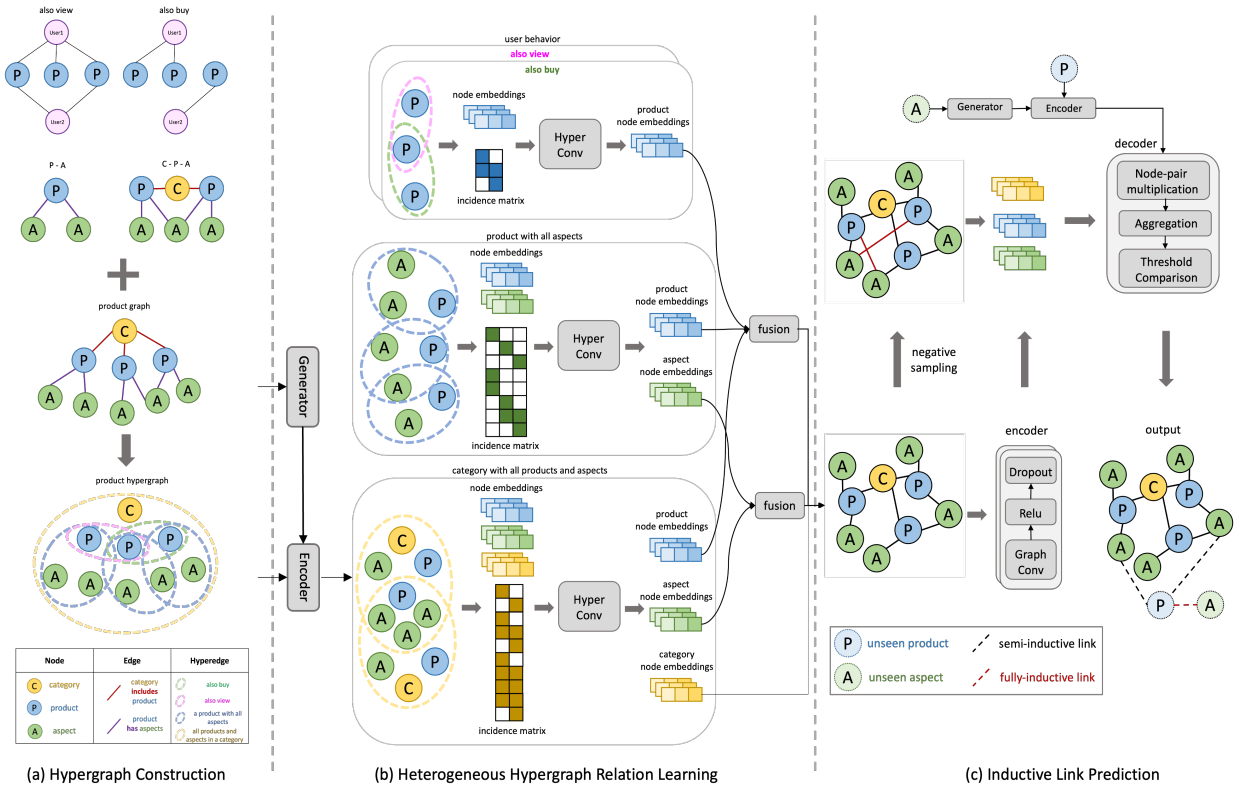


Figure 6.2: Overall framework of our proposed model HyperPAVE. The framework includes three key components: (a) Hypergraph Construction (b) Heterogeneous Hypergraph Relation Learning and (c) Inductive Link Prediction.

Figure 6.2 shows our proposed framework HyperPAVE with three main components: a) hypergraph construction, b) heterogeneous hypergraph relation learning, and c) inductive link prediction. We introduce each component in detail below.

### Heterogeneous Hypergraph Construction

As shown in Figure 6.2 (a), there are three types of nodes: categories, products, and attribute values (aspects), and four types of hyperedges: 'also view', 'also buy', 'product with all aspects' and 'category with all products and aspects', which are constructed from two main data sources: user behavior information and product inventory information as:

*1) User Behavior Data.* User behaviors have multiple types related to item-to-item relationships: people who bought X also bought Y ('also buy') and people who viewed X also viewed Y ('also view'). To well handle different user behaviors, we construct two types of hyperedges $\mathcal{T}_e^u = \{\mathcal{E}^V, \mathcal{E}^B\}$, where $\mathcal{E}^V$ represents 'also view' and $\mathcal{E}^B$ represents 'also buy'. For example, given the record of user1 in 'also view' graph shown in Figure 6.2 (a), we construct a hyperedge $\mathcal{E}_i^V = \{p_1, p_2, \cdots, p_n\} \in \mathcal{E}^V$ to model the interactions between users and products. That is, each hyperedge in $\mathcal{E}^V$ corresponds to one user. These hyperedges are homogeneous because all nodes represent products.

*2) Product Inventory Data.* Product inventory data refers to the existing product information records, including category, product, attribute values, etc. We construct hyperedges $\mathcal{E}^P$ to connect all attribute values to one product (P-A) and hyperedges $\mathcal{E}^C$ to connect all product information to one sub-category (C-P-A). For example, given a product $p_i$, we construct a hyperedge $\mathcal{E}_i^P = \{p_i, a_1, a_2, \cdots, a_n\} \in \mathcal{E}^P$ to indicate the relationships between product and its attribute values. These heterogeneous hyperedges record the non-binary relations among categories, products, and attribute values. To summarize it, we obtain hyperedge sets as:

$$\mathcal{T}_e = \{\mathcal{E}^V, \mathcal{E}^B, \mathcal{E}^P, \mathcal{E}^C\} \tag{6.3}$$

### Heterogeneous Hypergraph Relation Learning

**Embedding Module.** As shown in Figure 6.2 (b), a heterogeneous hypergraph encoder first initializes the node embeddings. Since the attribute values (aspects) may lose contextual information due to the simple format, GPT-2 [162] is adopted as the text generator to generate more detailed descriptions for attribute values. For example, the attribute value: 'connectivity: wireless' can be elaborated to a more detailed explanation: 'connectivity is wireless communication between the user's device, which has an independent, physical signal to the user'. We then adopt a pre-trained language model BERT [35] as all nodes' input encoder to generate the initial contextual representation. For the product node, we construct a string [CLS;$t$;SEP;$d$] by concatenating the product title and description as the input, where CLS and SEP are special tokens. The initial output representation for the category node $c_i$, product node $p_i$ and aspect node $a_i$ can be formulated as follows:

$$h_{v_{c_i}} = tanh(W \cdot f_{\varnothing}(c_i) + b) \tag{6.4}$$

$$h_{v_{p_i}} = tanh(W \cdot f_{\varnothing}(t_i, d_i) + b) \tag{6.5}$$

$$h_{v_{a_i}} = tanh(W \cdot f_{\varnothing}(g_{\varnothing}(a_i)) + b) \tag{6.6}$$

where $f_{\varnothing}$ is BERT encoder, $g_{\varnothing}$ is GPT-2 generator, $c$ is category, $t$ is product title, $d$ is product description, $a$ is 'attribute value', $W$ and $b$ are trainable weights and bias. To simplify the notations, we use $h_{v_i}$ to denote the initial feature embeddings of all different nodes.

**Message Passing Module.** To support representation learning on the constructed heterogeneous hypergraphs in the previous step, we design a heterogeneous hypergraph relation learning module (shown in Figure 6.2 (b) in HyperPAVE to explore the complex higher-order relationships based on many-to-many node message passing in the product graph by taking full advantage of the structure information in Figure 6.2 (a). HyperPAVE learns node representations with two different aggregation functions:

$$h_{v_i}^{l} = AGGR_{edge}^{l}\left(h_{v_i}^{l-1}, \left\{h_{e_j}^{l} | \forall_{e_j} \in \mathcal{E}_i\right\}\right) \tag{6.7}$$

$$h_{e_j}^{l} = AGGR_{node}^{l}\left(\left\{h_{v_k}^{l-1} | \forall_{v_k} \in e_j\right\}\right) \tag{6.8}$$

where $AGGR$ is the aggregation function, $\mathcal{E}_i$ is the hyperedge sets connected to node $v_i$ and $h_{e_j}^{l}$ is the representation of hyperedge $e_j$ in layer $l$. Since not all the nodes in a hyperedge will contribute equally, the message passing is calculated from nodes to hyperedges:

$$\alpha_{v_i}^{e_i} = \frac{exp(LeakyReLU(w_1^T \cdot h_{v_i}^{l-1}))}{\sum_{v \in V_{e_i}} exp(LeakyReLU(w_1^T \cdot h_v^{l-1}))} \tag{6.9}$$

$$h_{e_i}^{l} = ||_{n=1}^{N} \sigma(\sum_{v \in V_{e_i}} \alpha_v^{e_i} \cdot h_v^{l-1}) \tag{6.10}$$

where $\alpha_{v_i}^{e_i}$ is the weight factor of node $v_i$ to hyperedge $e_i$, $V_{e_i}$ is the node set of hyperedges $e_i$, $w_1^T$ is a trainable attention parameter, $||$ denotes concatenation with $N$ heads, and $\sigma$ is a non-linear function. $h_{e_i}^{l}$ is the $l^{th}$ layer of hyperedge representation. Similarly, the message passing from hyperedges to nodes is calculated as:

$$\alpha_{e_i}^{v_i} = \frac{exp(LeakyReLU(w_2^T \cdot (h_{v_i}^{l-1} || h_{e_i}^{l-1})))}{\sum_{e \in \mathcal{E}_{v_i}} exp(LeakyReLU(w_1^T \cdot (h_{v_i}^{l-1} || h_e^{l-1})))} \tag{6.11}$$

$$h_{v_i}^{l} = ||_{n=1}^{N} \sigma(\sum_{e \in \mathcal{E}_{v_i}} \alpha_e^{v_i} \cdot h_e^{l-1}) \tag{6.12}$$

where $\alpha_{e_i}^{v_i}$ is the weight factor of hyperedge $e_i$ to node $v_i$, $\mathcal{E}_{v_i}$ is the connected hyperedge set of node $v_i$. $w_2^T$ is a trainable attention parameter and $h_{v_i}^{l}$ is the $l^{th}$ layer of node representation, which includes the information from the hyperedge $\mathcal{E}$.

**Fusion Module** Instead of directly adding a readout layer and a linear prediction layer after obtaining the $L$ layers node representations [229], we argue that different types of hyperedges from $\mathcal{T}_e$ have different importance to the final node representations. Thus, we propose fusion modules to fuse node representations learned from different hypergraphs constructed in Sec. 6.3.3. The updated node representations for product node $\hat{h_{v_{p_i}}}$ and aspect node $\hat{h_{v_{a_i}}}$ are:

$$\hat{h_{v_{p_i}}} = \alpha \cdot h_{v_{p_i}}^{\mathcal{E}^{\mathcal{P}}} + \beta \cdot h_{v_{p_i}}^{\mathcal{E}^{\mathcal{C}}} + (1-\alpha-\beta)(\gamma \cdot h_{v_{p_i}}^{\mathcal{E}^{\mathcal{V}}} + (1-\gamma) \cdot h_{v_{p_i}}^{\mathcal{E}^{\mathcal{B}}}) \tag{6.13}$$

$$\hat{h_{v_{a_i}}} = \delta \cdot h_{v_{a_i}}^{\mathcal{E}^{\mathcal{P}}} + (1-\delta) \cdot h_{v_{a_i}}^{\mathcal{E}^{\mathcal{C}}} \tag{6.14}$$

where $h_{v_{p_i}}^{\mathcal{E}^{\mathcal{P}}}$, $\cdot h_{v_{p_i}}^{\mathcal{E}^{\mathcal{C}}}$, $h_{v_{p_i}}^{\mathcal{E}^{\mathcal{V}}}$, $h_{v_{p_i}}^{\mathcal{E}^{\mathcal{B}}}$ are product node representations and $h_{v_{a_i}}^{\mathcal{E}^{\mathcal{P}}}$, $h_{v_{a_i}}^{\mathcal{E}^{\mathcal{C}}}$ are aspect node representations from different hyperedges in Equ. 6.3, respectively. $\alpha$, $\beta$, $\gamma$, and $\delta$ are weights learnt from the validation sets. They are different for different categories of the dataset. These weights are also explored and studied in Sec. 6.4.2. After the above fusion steps, the node embeddings contain the features from neighbors defined by different hyperedges $\mathcal{T}_e$, which can well capture the high-order relations communicated among different types of nodes and hyperedges.

### Inductive Link Prediction

After heterogeneous hypergraph relation learning, each node includes the higher-order features related to user behavior and product inventory information. Then, all the nodes go through $L$ GNN layers to compute the final node representations. After generating the final embeddings of $\tilde{h_{v_p}}$ and $\tilde{h_{v_a}}$, the likelihood of the link between product $p$ and aspect $a$ is measured by the cosine similarity to decide the possibility $\hat{R}_{ij}$ of whether product $p_i$ will have the aspect $a_j$:

$$f_{score}((\tilde{h_{v_p}})_i, (\tilde{h_{v_a}})_j) = \frac{(\tilde{h_{v_p}})_i \cdot (\tilde{h_{v_a}})_j}{\left\|(\tilde{h_{v_p}})_i\right\| \left\|(\tilde{h_{v_a}})_j\right\|} \tag{6.15}$$

We use the negative sampling strategy introduced in Sec. 6.3.2 to train HyperPAVE and employ a binary cross entropy loss to optimize our model:

$$\mathcal{L} = \sum_{p_i \in P, a_i \in A} R_{ij} log\hat{R}_{ij} + (1-R_{ij})(1-log\hat{R}_{ij}) \tag{6.16}$$

Note that HyperPAVE follows the multi-label zero-shot settings in Sec. 6.3.2 to eliminate the mandatory access of testing node features during training, making the model access the inductive inference ability. For unseen attribute values (aspects) and products, we can directly feed their corresponding contextual node embeddings by fine-tuned BERT encoder to HyperPAVE instead of representing product and aspect nodes with one-hot vectors. Then,

Table 6.1: Dataset statistics over ten categories. The number of hyperedges is reported in the format of: #nodes / #hyperedges.

| Category | Number of Nodes | | | Number of Edges | | |
|---|---|---|---|---|---|---|
| | #C | #P | #A | #CP | #PA | Density |
| Arts | 980 | 11,625 | 2,184 | 50,652 | 28,932 | $7.2\times10^{-4}$ |
| Books | 410 | 16,220 | 255 | 48.271 | 23,438 | $5.03\times10^{-4}$ |
| Cellphones | 145 | 8,499 | 1,484 | 27,620 | 20,329 | $9.35\times10^{-4}$ |
| Giftcards | 5 | 131 | 11 | 378 | 311 | 0.06 |
| Grocery | 742 | 18,315 | 4,686 | 75,362 | 47,745 | $4.37\times10^{-4}$ |
| Industrial | 433 | 3002 | 1573 | 12,429 | 8,453 | $1.67\times10^{-3}$ |
| Pet | 508 | 14,299 | 2,575 | 64,947 | 46,370 | $7.34\times10^{-4}$ |
| Software | 303 | 254 | 98 | 1,182 | 607 | $8.35\times10^{-3}$ |
| Tools | 975 | 34,076 | 7,538 | 143,683 | 101,475 | $2.7\times10^{-4}$ |
| Videogames | 910 | 731 | 353 | 4,446 | 2,152 | $3.32\times10^{-3}$ |

we only conduct message-passing and compute the probability of connections between the product node and the aspect node. Hence, we can handle the newly added products and attribute values in an inductive way instead of retraining the model.

## 6.4 Experiments

### 6.4.1 Experimental Setup

#### Dataset

We evaluate our model over ten different categories (Arts, Books, Cellphones, etc) of a public dataset MAVE [235], which is a large e-Commerce dataset derived from Amazon Review Dataset [151]. To simulate the zero-shot situation, we reconstruct the dataset into multi-label zero-shot learning settings followed by Sec. 6.3.2, where there is no overlap of products and attribute values between the training set and validation/testing set. Note that each time we train the model, the dataset will be randomly re-splitted for the zero-shot setting, so we report the whole data statistics in Table 6.1.

Table 6.2 reports an example of dataset statistics in training, validation, and testing sets, where $\#P$, $\#A$, and $\#PA$ denotes the number of product nodes, the number of aspect nodes and the number of product to aspect edges, respectively. The last column Ave $\#A/p$ indicates the average number of attribute value pairs for each product. Because training, validation, and testing sets for the multi-label zero-shot setting are randomly generated for each run of the experiment, there exist different dataset statistics.

| Category | Number of Hyperedges | | | |
|---|---|---|---|---|
| | P-$P_{\text{also view}}$ | P-$P_{\text{also buy}}$ | P-A | C-P-A |
| Arts | 970/624 | 1,448/1,248 | 13,809/11,643 | 14,789/979 |
| Books | 1,247/1,433 | 2,432/2,550 | 16,475/16,222 | 16,885/409 |
| Cellphones | 366/362 | 171/160 | 9,983/8,507 | 10,128/144 |
| Giftcards | 17/20 | 19/32 | 142/130 | 147/1 |
| Grocery | 3,162/2,431 | 3,392/3,314 | 23,001/4,686 | 23,743/741 |
| Industrial | 152/106 | 210/205 | 4,539/3,063 | 5,008/432 |
| Pet | 1,614/1,670 | 820/600 | 16,874/14,675 | 17,382/507 |
| Software | 19/20 | 2/1 | 352/287 | 655/302 |
| Tools | 3,176/2,648 | 1,998/1,704 | 41,614/34,705 | 42,589/974 |
| Videogames | 113/139 | 14/9 | 1,084/752 | 1,994/909 |

Table 6.2: Example of zero-shot dataset statistics in training, validation and testing sets, respectively.

| Category | Training | | | Validation | | | Testing | | | All |
|---|---|---|---|---|---|---|---|---|---|---|
| | #P | #A | #PA | #P | #A | #PA | #P | #A | #PA | Ave #A/P |
| Arts | 10,250 | 1,796 | 8,400 | 3 | 6 | 6 | 15 | 23 | 30 | 2.48 |
| Books | 9,310 | 158 | 5,210 | 4 | 3 | 8 | 413 | 54 | 852 | 1.44 |
| Cellphones | 6,772 | 1,149 | 5,187 | 91 | 109 | 192 | 157 | 175 | 332 | 2.38 |
| Giftcards | 84 | 8 | 74 | 8 | 2 | 16 | 11 | 3 | 9 | 2.37 |
| Grocery | 15,834 | 3,945 | 13,933 | 8 | 16 | 16 | 18 | 33 | 36 | 2.56 |
| Industrial | 2,644 | 1,264 | 2,381 | 16 | 27 | 33 | 8 | 14 | 17 | 2.76 |
| Pet | 12,878 | 2,193 | 13,187 | 24 | 42 | 48 | 73 | 117 | 150 | 3.16 |
| Software | 187 | 87 | 152 | 2 | 4 | 4 | 8 | 14 | 16 | 2.11 |
| Tools | 30,236 | 6,210 | 29,759 | 14 | 24 | 28 | 58 | 97 | 120 | 2.92 |
| Videogames | 559 | 240 | 477 | 35 | 45 | 75 | 57 | 67 | 128 | 2.86 |

### Evaluation Metrics

Following other AVE tasks in the multi-label zero-shot setting [188], we choose to report macro-F1 and mAP (mean Average Precision) compared with classification and generation-based models in the main results as F1 score is the balance of both precision and recall. In Sec. 4.4.4 ablation study, we also report AUC (Area Under Curve), Hits@K, NDCG@K (Normalized Discounted Cumulative Gain), and MRR (Mean Reciprocal Ran), which are widely used metrics in graph-based recommendation tasks [46, 74, 114]. We also report training time to evaluate the efficiency in Sec. 6.4.2 efficiency study.

### Baselines

We compare our proposed model HyeprPAVE with the following baselines in the zero-shot setting:

- Classification-based Models: Original classification-based models do not have any zero-shot abilities. We follow the baseline **BERT-MLC** in [188], then we add synthetic data for unseen classes (attribute values) following [29]. In this way, the zero-shot learning problem is translated into a supervised learning problem.
- Generation-based Models: Following generative models in zero-shot AVE task [188], we implement and fine-tune two text-to-text transformer-based encoder decoder architecture models: **BART** [102] and $\mathbf{T5_{small}}$ [163], to generate unseen attribute values directly.
- Graph-based Models [1]: As inductive graph can predict unseen nodes, we compare HyperPAVE with three heterogeneous GNNs: **HGCN** [164], **HAN** [209], **HGT** [82], and two representative hypergraph networks: **hyperGCN** [230], **HGNN+** [54].

## 6.4.2 Parameter Settings

We randomly select unseen attribute value pairs with unseen products following the sampling rule in Sec. 6.3.2. For the hyperparameter and configuration of HyperPAVE, we implement HyperPAVE in PyTorch and optimize it with AdamW optimizer. We train HyperPAVE and all baselines on the training set and we use a validation set to select the optimal hyper-parameter settings, and finally report the performance on the test set. We follow the early stopping strategy when selecting the model for testing. For all methods, we run 10 times with different random seeds and report the average results with standard deviation. Our proposed model HyperPAVE achieves its best performance with the following setup. The nodes' features are initialized by a BERT encoder with a 768-dimension size. The max length

[1] Implemented on DHG: https://deephypergraph.com/

Table 6.3: Experimental Results F1 / mAP (%) of multi-label zero-shot learning over ten categories on MAVE. The results are reported as mean over ten times of experiments. The best results are in bold.

| | Arts | Books | Cellphones | Giftcards | Grocery |
|---|---|---|---|---|---|
| BERT-MLC [22] | 24.11 / 10.31 | 36.72 / 27.17 | 22.92 / 28.67 | 36.54 / 41.15 | 19.74 / 12.07 |
| Bart [102] | 27.88 / 23.16 | 38.82 / 44.90 | 32.71 / 24.54 | 15.73 / 8.75 | 10.80 / 6.95 |
| $T5_{small}$ [163] | 30.85 / 23.16 | 36.17 / 42.60 | 30.95 / 24.27 | 10.14 / 8.08 | 23.53 / 17.32 |
| HGCN [164] | 16.87 / 25.30 | 39.39 / 37.40 | 17.23 / 14.67 | 30.92 / 45.42 | 25.60 / 39.77 |
| HAN [209] | 14.26 / 26.42 | 43.73 / 49.48 | 22.49 / 33.69 | 42.47 / 54.05 | 17.23 / 34.67 |
| HGT [82] | 30.81 / 38.53 | 48.06 / 41.67 | 14.53 / 23.73 | 42.30 / 42.39 | 27.30 / 40.76 |
| HGNN+ [54] | 27.90 / 36.91 | 46.79 / **58.33** | 32.10 / **36.40** | 37.18 / 57.20 | 32.40 / 38.60 |
| HyperGCN [230] | 20.20 / 38.45 | 48.97 / 45.18 | 20.90 / 26.00 | **52.74** / 45.97 | **35.90** / 42.20 |
| **HyperPAVE** | **43.33 / 40.99** | **49.75** / 56.45 | **39.01** / 35.81 | 52.34 / **65.03** | 33.43 / **42.71** |
| | Industrial | Pet | Software | Tools | Videogames |
| BERT-MLC [22] | 10.94 / 6.69 | 18.14 / 12.08 | 27.76 / 25.37 | 20.43 / 18.41 | 11.86 / 9.66 |
| BART [102] | 10.78 / 7.84 | 12.50 / 10.42 | 22.50 / 20.00 | 11.11 / 6.25 | 23.57 / 20.02 |
| $T5_{small}$ [163] | 15.81 / 15.35 | 25.28 / 25.72 | 26.19 / 24.60 | **37.78** / 22.46 | 14.41 / 9.90 |
| HGCN [164] | 10.67 / 14.60 | 17.62 / 24.63 | 19.29 / 30.97 | 18.07 / 39.32 | 8.78 / 13.61 |
| HAN [209] | 15.35 / 30.45 | 16.82 / 23.33 | 28.24 / 29.03 | 19.78 / 41.40 | 9.68 / 16.29 |
| HGT [82] | 21.09 / 23.20 | 18.02 / 23.66 | 30.15 / 27.16 | 13.61 / 35.23 | 14.75 / 19.97 |
| HGNN+ [54] | 25.90 / 28.60 | 27.60 / 35.58 | 39.90 / 28.76 | 31.00 / 42.20 | 10.35 / 17.21 |
| HyperGCN [230] | **29.20** / 33.20 | 22.20 / 31.37 | 42.10 / 38.70 | 31.10 / 44.05 | 10.90 / 15.30 |
| **HyperPAVE** | 27.70 / **33.29** | **28.45 / 38.46** | **47.62 / 51.64** | 34.00 / **47.83** | **25.31 / 21.19** |

for category, product, and attribute values are 32, 512, and 32, respectively. The initial learning rate is selected via grid search within the range of $\{5e-1, 5e-3, 5e-4, 5e-5\}$ with 1e-6 weight decay for minimizing the loss. The hidden sizes for convolution layers are 768 in both HyperConv and GraphConv. The activation function is ReLU. The dropout rate is 0.5 and the batch size is 4. We set the number of neighbors to 20 and the negative sampling rate is 2.0. For the fusion module, the weights of the product node embeddings from hyperedges of 'also buy', 'also view', 'products with all aspects', and 'category with all products and aspects' are dynamically changed for different categories. Experiments are conducted in Sec. 6.4.2 to explore the weights in these fusion modules.

### Main Results

From the results shown in Table 6.3 and data statistics shown in Table 6.1, we observe that:

(1) The classification-based model generally performs worst among all models. BERT-MLC, which uses synthetic data for zero-shot prediction, only has competitive performance to generation-based models when the class number (#A) is small. We conjecture that as the number of classes grows, BERT-MLC needs to make distinctions among more classes, mak-

ing it harder to find clearer decision boundaries. The average micro F1 of BART and T5 $_{\text{small}}$ across all ten categories is worse than T5 $_{\text{base}}$ in [188]. This is because T5 $_{\text{base}}$ is pre-trained over 220 million parameters whereas T5 $_{\text{small}}$ has only 60 million parameters. Generation-based models perform much better than classification-based models in most cases. BART and T5 $_{\text{small}}$ show different performances over different categories. They can achieve similar performance with HyperPAVE when the dataset size is large enough. (2) Combining inductive graph-based models with LLM encoders can perform zero-shot prediction and achieve competitive performance with generative models. This inspires us that instead of fine-tuning the popular generative models [177, 178, 188, 256] to extract attribute values, the inductive graph for link prediction can also be explored for zero-shot prediction. Besides, using the attention mechanism shows better performance than using fixed and uniform weights for aggregation. This is probably because assigning different weights to neighboring nodes can capture varying levels of influence. (3) Compared with graph-based baselines, adding complex structured data to capture higher-order relationships demonstrates significant performance improvement over all ten categories. This is probably because hyperedges can model relationships that go beyond pairwise connections, resulting in more semantic node representations. Besides, our proposed model HyperPAVE achieves the best performance among all models in most categories, indicating that our proposed hypergraph construction from both user behavior data and product inventory data is important and worth recording and exploring. The effectiveness of different hyperedges is studied in Sec. 6.4.2.

**Ablation Study**

To evaluate the performance of each component in HyperPAVE, we conduct an ablation study over three categories in the zero-shot setting. Table 6.4 shows the performance of each component in HyperPAVE. We have the following observations based on Table 6.4: (1) Adding node features can significantly improve the performance. We perform a model 'nodeID', which doesn't use any pre-trained encoder for providing node features. The model 'nodeID' uses a simple embedding-lookup encoder, mapping each node to a unique low-dimensional vector. We can observe that among all models, 'nodeID' shows the worst performance. After adding node features, such as BERT or fine-tuned BERT, the performance increases significantly. We think that this is because, for link prediction in the zero-shot setting, pre-trained embeddings provide richer and more semantically meaningful representations for node features in graphs than simple one-hot encoding. (2) Fine-tuning the pre-trained encoders for node features results in a big performance improvement when the dataset (graph) is large enough. This is reasonable because a larger dataset (graph with more nodes) provides more diverse and representative data, enabling better generalization for unseen nodes in the zero-shot setting. However, as shown in Sec. 6.4.2, fine-tuning the pre-trained encoder may result in more time for model training. A balance of model performance and efficiency needs to be considered for different tasks/situations. (3) We explore the importance of different hy-

Table 6.4: Ablation study over HyperPAVE components in the zero-shot setting across ten categories on MAVE dataset.

| | F1 | mAP | AUC | MRR | NDCG | Hits@5 | Hits@10 | Hits@100 |
|---|---|---|---|---|---|---|---|---|
| | Arts | | | | | | | |
| nodeID | 1.35 | 10.73 | 92.03 | 0.86 | 21.92 | 15.00 | 28.33 | 61.67 |
| BERT | 8.23 | 26.30 | 92.69 | 11.48 | 40.27 | 35.43 | 52.86 | 82.43 |
| BERT (Fine-tuned) | 19.87 | 34.77 | **99.84** | 13.16 | 53.84 | 75.00 | 75.00 | 100.00 |
| Hyper (Product) | 30.93 | 42.33 | 99.06 | 22.95 | 57.84 | 57.50 | 75.00 | 93.75 |
| Hyper (Behavior) | 37.03 | 42.39 | 98.35 | 28.51 | 60.47 | **83.33** | 100.00 | 100.00 |
| **HyperPAVE** | **43.33** | **40.99** | 99.22 | **47.52** | **64.87** | **75.00** | 82.50 | **100.00** |
| | Books | | | | | | | |
| nodeID | 11.54 | 28.52 | 95.31 | 6.64 | 48.41 | 35.26 | 53.85 | 99.42 |
| BERT | 23.87 | 38.63 | 97.07 | 11.39 | 57.31 | 47.05 | 63.59 | 100.00 |
| BERT (Fine-tuned) | 28.28 | 40.32 | 97.87 | 14.44 | 58.89 | 50.90 | 78.33 | 100.00 |
| Hyper (Product) | 30.44 | 40.65 | 98.03 | 14.23 | 59.49 | 49.36 | 80.51 | 100.00 |
| Hyper (Behavior) | 34.46 | 35.93 | **98.40** | 19.37 | 54.23 | 63.67 | 93.67 | 100.00 |
| **HyperPAVE** | **49.75** | **56.45** | 96.47 | **32.99** | **69.35** | **85.27** | **94.04** | **100.00** |
| | Cellphones | | | | | | | |
| nodeID | 19.75 | 22.88 | 97.72 | 10.65 | 38.33 | 38.33 | 57.22 | 80.00 |
| BERT | 24.21 | 26.15 | 97.60 | 17.02 | 40.97 | 41.11 | 70.05 | 86.11 |
| BERT (Fine-tuned) | 22.77 | 26.80 | 98.09 | 16.50 | 43.03 | 50.00 | 75.00 | 92.50 |
| Hyper (Product) | 32.27 | 33.32 | 98.94 | 22.26 | **54.17** | 70.25 | **90.00** | 100.00 |
| Hyper (Behavior) | 28.32 | 33.88 | **99.63** | 22.57 | 47.81 | 52.50 | 61.67 | 97.50 |
| **HyperPAVE** | **39.91** | **35.81** | 99.22 | **23.54** | 52.88 | **72.50** | 75.00 | **100.00** |
| | Giftcards | | | | | | | |
| nodeID | 6.67 | 22.35 | 41.94 | 18.18 | 42.92 | 25.00 | 97.50 | 100.00 |
| BERT | 26.41 | 44.79 | 71.94 | 24.15 | 62.47 | 75.00 | 100.00 | 100.00 |
| BERT (Fine-tuned) | 34.43 | 41.67 | 71.53 | 23.17 | 59.57 | 67.50 | 100.00 | 100.00 |
| Hyper (Product) | 39.77 | 45.74 | 84.55 | 35.65 | 61.83 | 100.00 | 100.00 | 100.00 |
| Hyper (Behavior) | 45.43 | 60.50 | 77.92 | 29.13 | 73.02 | 71.00 | 100.00 | 100.00 |
| **HyperPAVE** | **52.34** | **65.03** | **90.08** | **44.56** | **75.07** | **100.00** | **100.00** | **100.00** |
| | Pets | | | | | | | |
| nodeID | 6.95 | 13.46 | 98.51 | 7.47 | 42.17 | 30.33 | 50.00 | 96.15 |
| BERT | 9.93 | 19.93 | 99.73 | 6.71 | 41.16 | 31.67 | 65.00 | 100.00 |
| BERT (Fine-tuned) | 12.12 | 19.99 | 99.39 | 10.79 | 40.52 | 25.00 | 56.67 | 100.00 |
| Hyper (Product) | 17.58 | 37.40 | 99.03 | 16.45 | 45.67 | 41.67 | **71.67** | 98.33 |
| Hyper (Behavior) | 18.66 | 24.71 | 99.16 | 19.01 | 42.38 | 36.07 | 65.00 | 100.00 |
| **HyperPAVE** | **28.45** | **38.46** | **99.82** | **29.92** | **61.55** | **56.67** | 67.77 | **100.00** |

| | F1 | mAP | AUC | MRR | NDCG | Hit@5 | Hits@10 | Hits@100 |
|---|---|---|---|---|---|---|---|---|
| | Grocery | | | | | | | |
| nodeID | 6.50 | 23.31 | 95.48 | 15.33 | 21.98 | 22.50 | 35.00 | 65.00 |
| BERT | 14.65 | 22.85 | 96.18 | 15.80 | 22.55 | 30.10 | 35.10 | 75.00 |
| BERT (Fine-tuned) | 19.42 | 25.84 | 99.20 | 17.78 | 27.93 | 25.00 | 35.50 | **87.50** |
| Hyper (Product) | 22.41 | 32.41 | 99.48 | 18.82 | 35.64 | 33.33 | 35.50 | 66.67 |
| Hyper (Behavior) | 29.20 | 32.85 | 98.34 | 14.41 | 37.66 | 35.05 | 50.00 | 70.00 |
| **HyperPAVE** | **33.43** | **42.71** | **99.56** | **22.52** | **52.64** | **50.00** | **50.00** | 75.50 |
| | Industrial | | | | | | | |
| nodeID | 10.40 | 16.44 | 93.16 | 2.59 | 30.07 | 28.75 | 35.00 | 68.75 |
| BERT | 1.48 | 5.37 | 89.75 | 0.66 | 13.58 | 8.13 | 11.87 | 55.63 |
| BERT (Fine-tuned) | 14.06 | 18.82 | 99.05 | 4.99 | 41.11 | 25.00 | 50.00 | 100.00 |
| Hyper (Product) | 19.78 | 14.15 | 94.34 | 7.63 | 26.68 | 24.73 | 37.50 | 75.00 |
| Hyper (Behavior) | 15.70 | 31.42 | 96.57 | 7.19 | 45.26 | 41.25 | 55.00 | 87.50 |
| **HyperPAVE** | **27.70** | **33.29** | **99.71** | **16.10** | **54.08** | **52.50** | **80.00** | **100.00** |
| | Software | | | | | | | |
| nodeID | 1.97 | 18.11 | 76.27 | 4.39 | 30.12 | 23.75 | 62.50 | 100.00 |
| BERT | 7.38 | 14.10 | 74.89 | 6.38 | 34.19 | 26.70 | 36.25 | 100.00 |
| BERT (Fine-tuned) | 11.78 | 15.29 | 76.70 | 6.75 | 36.52 | 23.75 | 37.50 | 100.00 |
| Hyper (Product) | 35.88 | 40.72 | **84.40** | 21.25 | 59.51 | 46.25 | **63.75** | 100.00 |
| Hyper (Behavior) | 12.22 | 36.33 | 81.25 | 6.09 | 34.19 | 25.00 | 38.75 | 100.00 |
| **HyperPAVE** | **47.62** | **51.64** | 77.80 | **26.66** | **63.48** | **61.25** | 62.50 | **100.00** |
| | Tools | | | | | | | |
| nodeID | 8.90 | 17.00 | 97.91 | 2.36 | 22.27 | 50.00 | 50.00 | 50.00 |
| BERT | 14.53 | 18.51 | 96.21 | 6.51 | 21.30 | 48.50 | 52.05 | 80.00 |
| BERT (Fine-tuned) | 21.33 | 23.85 | 99.19 | 6.81 | 26.88 | 49.15 | 55.70 | **87.07** |
| Hyper (Product) | 32.86 | 29.20 | 98.27 | 12.26 | 43.96 | 49.53 | 65.00 | 83.87 |
| Hyper (Behavior) | 31.43 | 25.13 | **99.30** | 11.51 | 28.11 | 50.06 | 58.20 | 86.40 |
| **HyperPAVE** | **34.00** | **47.83** | 98.00 | **12.93** | **59.05** | **52.00** | **65.37** | 84.72 |
| | Videogames | | | | | | | |
| nodeID | 3.25 | 7.31 | 79.00 | 1.49 | 17.27 | 10.00 | 20.00 | 70.00 |
| BERT | 6.67 | 10.25 | 85.83 | 3.01 | **33.30** | 30.05 | 43.50 | 100.00 |
| BERT (Fine-tuned) | 12.87 | 11.44 | 76.84 | 4.21 | 25.26 | 15.71 | 37.86 | 73.70 |
| Hyper (Product) | 20.00 | 16.45 | 91.51 | 8.76 | 28.61 | 45.00 | 50.00 | **100.00** |
| Hyper (Behavior) | 16.83 | 12.38 | 86.73 | 7.33 | 27.28 | 15.00 | 40.71 | 80.71 |
| **HyperPAVE** | **25.31** | **21.19** | 84.32 | **9.31** | 23.99 | **50.00** | **50.00** | 85.71 |

Table 6.5: Comparison of computational efficiency. The batch size is set to 4.

| Model | Memory Consumption | Model Parameters |
|---|---|---|
| Classification-based | 5037MB | 110M |
| Generation-based | 8305MB / 5831MB | 140M / 60M |
| Graph-based (ours) | 1405MB / 1915 MB | 5M / 115M |

pergraphs. We find out that adding different hyperedges built from user behavior data or product inventory data results in a significant performance improvement. We conjecture that this is because different hyperedges capture more complex higher-order information than the original binary-relation graph. For example, hyperedge 'P-$P_{also\_view}$' built from user behavior data includes information on products with potential similar attributes because users may probably view similar products at the same time for their needs. Hyperedge 'C-P-A', built from product inventory data, aggregates all products and aspects in the same sub-category. Attribute values such as 'Chew Type: Bones' may only happen in a sub-category of 'Dog Treats' instead of 'Cat Food'. By using hyperedges, more complex relations can be included in the representation for each single node.

**Efficiency Study**

Table 6.5 presents the GPU computational cost and model parameter comparison between classification-based (BERT-MLC), generation-based (BART/$T5_{small}$) and graph-based (nodeID/HyperPAVE) models on Arts category of MAVE. Different categories (different sizes of graphs) may result in a slight difference. From the reported results, we can find that compared with classification or generation-based models, our proposed graph-based model HyperPAVE, has a significant computational advantage in terms of memory consumption. The main reason is that the zero-shot ability of generative LLMs is based on their extensive pretraining and understanding of the diverse data. When fine-tuning these LLMs, large quantities of model parameters need to be updated, resulting in a huge GPU memory consumption cost. However, the zero-shot ability of HyperPAVE results from the inductive inference that can generalize to unseen product and aspect nodes without retraining the whole model. The inductive HyperPAVE divides the hypergraphs into batches and only consumes per-batch memory when training. Note that for the classification model BERT-MLC, preprocessing steps for generating synthetic data from generation models are required to predict unseen aspects. We have not counted the computational cost for these preprocessing steps.

To evaluate the computation time of our graph-based model and other classification-based and generation-based models, we record the model training time for one epoch in seconds across the ten categories on MAVE as shown in Figure 6.3. All models use the same input max_length and batch size for training. From Figure 6.3, we observe that graph-based models show better model training efficiency. Compared with other graph-based models

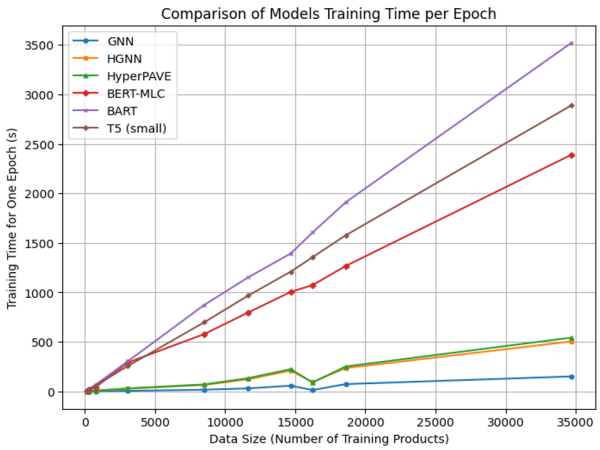


Figure 6.3: Time Efficiency Performance (GPU Time of Model Learning in Seconds for One Training Epoch).

(i.e. GNN, HGNN), HyperPAVE can achieve the best prediction performance as shown in Table 6.3 with only sacrificing a little more time for training as shown in Figure 6.3. The slopes of BART, T5, and BERT-MLE are much larger than graph-based models, indicating that much more time is needed for training or fine-tuning with the increase in dataset size when updating the model parameters. We also demonstrate the model training time for one epoch across the ten categories on MAVE in Table 6.6. All models use the same input max_length as 512 and batch size as 4. For different graph-based models, they show similar efficiency performance. Thus, we only demonstrate two representative graph-based models (GNN and HGNN) for training efficiency comparison.

Table 6.6: Model Training Time in One Epoch (second).

| Model | Gift. | Soft. | Video | Indus. | Cell. | Arts | Pet | Books | Groc. | Tools |
|---|---|---|---|---|---|---|---|---|---|---|
| BERT-MLC | 2.37 | 15.48 | 42.12 | 291.60 | 578.78 | 797.11 | 1004.53 | 1073.65 | 1266.68 | 2391.12 |
| BART | 3.66 | 24.48 | 66.60 | 304.56 | 873.36 | 1152.00 | 1292.95 | 1604.52 | 1910.52 | 3521.88 |
| $T5_{small}$ | 2.21 | 19.14 | 58.23 | 256.70 | 698.10 | 967.87 | 1209.27 | 1355.23 | 1576.67 | 2890.03 |
| GNN | 0.09 | 0.19 | 1.00 | 5.46 | 16.46 | 30.02 | 57.50 | 12.80 | 73.33 | 150.91 |
| HGNN | 0.72 | 1.60 | 6.28 | 27.59 | 64.24 | 122.06 | 209.28 | 94.52 | 235.20 | 504.61 |
| **Ours** | 0.90 | 1.66 | 6.71 | 30.06 | 70.18 | 133.43 | 224.40 | 89.22 | 251.14 | 543.07 |

### Parameter Sensitivity Analysis

The key hyperparameters of HyperPAVE are the weights of the different hyperedges. Thus, we explore the importance of different types of hyperedges in the category of Giftcards as shown in Figure 6.4.

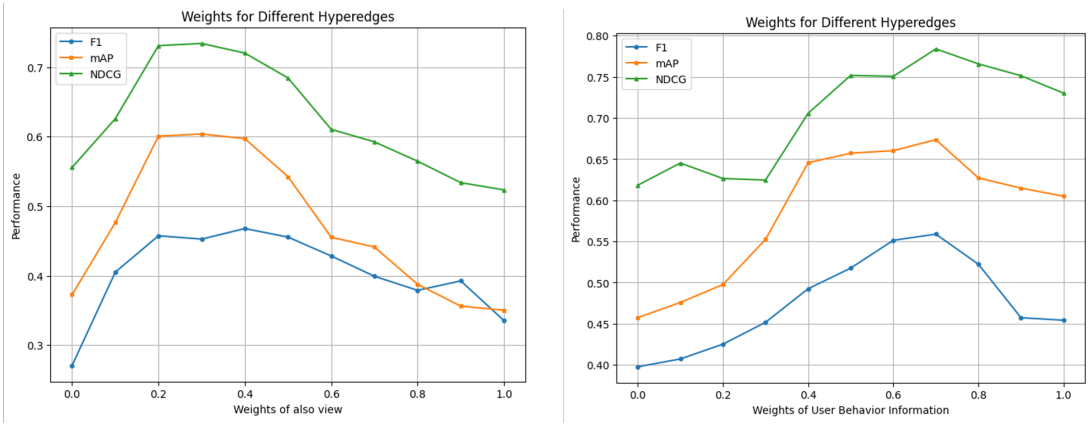


Figure 6.4: Effects on weights of different hyperedges on the category of giftcards.

The left figure explores the weights of 'P-$P_{\text{also view}}$' and 'P-$P_{\text{also buy}}$' hyperedges from user behavior information. The right figure explores the weights of user behavior hyperedges (P-P) and product inventory hyperedges ('P-A' and 'C-P-A'). From Figure 6.4, we observe that both 'P-$P_{\text{also view}}$' and 'P-$P_{\text{also buy}}$' contribute to the model's performance. The best weight for 'P-$P_{\text{also view}}$' falls in the [0.2, 0.5] interval, which means 'P-$P_{\text{also buy}}$' is slightly more important than 'P-$P_{\text{also view}}$'. This is probably because 'P-$P_{\text{also buy}}$' records users' history preference while 'P-$P_{\text{also view}}$' may include some noise such as accidental clicks. We can also observe from the right 6.4 that the best weight for user behavior data falls in the [0.6, 0.8] interval, indicating that user behavior is much more important than product inventory data. As shown in Table 6.1, the number of user behavior hyperedges is much smaller than the number of product inventory hyperedges ('P-A' and 'C-P-A'). But they show more importance in Figure 6.4, demonstrating that user behavior information is worth recording and exploring for extracting unseen attribute values for new products.

## 6.5 Summary

In this paper, we formulate the AVE task in a zero-shot learning scenario to identify unseen attribute values from new products with no corresponding labeled data available for training. We propose an inductive heterogeneous hypergraph (HyperPAVE) for multi-label zero-shot attribute value extraction. Specifically, the heterogeneous hypergraph captures the higher-order relationships among users and products, and the inductive mechanism infers the future connections between unseen nodes. Extensive experimental results on ten different categories across the public dataset MAVE demonstrate that our proposed model HyperPAVE outperforms other state-of-the-art classification-based and generation-based models. The ablation study validates the efficiency and effectiveness of different hypergraphs constructed from user behavior and product inventory data. We plan to explore the following directions in future work: (1) Including multimodal features (i.e. product images) as node attributes

to capture more semantic information from the products. (2) Building dynamic graphs by including timestamps to make the product graph adapt to the developing market.

# Chapter 7

# Conclusion and Future Work

## 7.1 Conclusion

In this dissertation, we answer the research question of *How can multi-source learning improve few-shot and zero-shot information extraction?* by studying and exploring various auxiliary information in few-shot and zero-shot learning settings for information extraction. For information extraction tasks, we mainly focus on general relation extraction and aspect (attribute-value) extraction tasks in the e-commerce application field. For few-shot information extraction, we propose a multi-modal model that leverages both textual and visual semantics to supplement the missing contexts in a single modality. Besides, we study the multi-label few-shot information extraction by proposing anchor-enhanced prototypical networks with a learnt dynamic threshold for multi-label inference.

For zero-shot information extraction, we first explore different side information including data statistics, keywords, labels and synonyms of labels, name entities and their hypernyms, etc. to enrich the information for zero-shot classes. Besides, we construct prompts based on an external knowledge graph to integrate semantic knowledge from seen classes with zero-shot classes, to transform the zero-shot learning to the supervised-learning task. Furthermore, we utilize semi-inductive link prediction of the heterogeneous hypergraph to predict zero-shot aspects via considering higher-order relations through auxiliary information of user behavior data and product inventory data.

## 7.2 Publications

A number of papers have been published or accepted during my Ph.D. program. A selected set of my first-author publications is shown in the list below.

- **J. Gong**, H. Eldardiry, *Multi-Label Zero-Shot Product Attribute-Value Extraction.* In Proceedings of the ACM Web Conference 2024 (WWW'24).

- **J. Gong**, H. Eldardiry, *Prompt-based Zero-shot Relation Extraction with Semantic Knowledge Augmentation.* In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).
- **J. Gong**, H. Eldardiry, *Few-Shot Relation Extraction with Hybrid Visual Evidence.* In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)
- **J. Gong**, W.Chen, H. Eldardiry, *Knowledge-Enhanced Multi-Label Few-Shot Product Attribute-Value Extraction.* In Proceedings of the 32nd ACM International Conference on Information & Knowledge Management (CIKM'23).
- **J. Gong**, H. Eldardiry, *Zero-shot Relation Classification from Side Information.* In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM'21).
- **J. Gong**, H. Eldardiry, *Multi-Stage Hybrid Attentive Networks for Knowledge-Driven Stock Movement Prediction.* Neural Information Processing: 28th International Conference (ICONIP'21).
- **J. Gong**, B. Paye, G. Kadlec, H. Eldardiry, *Predicting Stock Price Movement Using Financial News Sentiment.* 2021 International Conference on Engineering Applications of Neural Networks (EANN'21).
- **J. Gong***, M. Cheng*, C. Yuan, W. Ingram, E. Fox, H. Eldardiry, *VTechAGP: An Academic-to-General-Audience Text Paraphrase Dataset and Benchmark Models* (Under Review)
- **J. Gong**, M. Cheng, H. Shen, P. Vandenbussche, S. Khadivi, H. Eldardiry, Cross-Modal Zero-Shot Aspect Generation (Under Review)

## 7.3 Future Work

In the future, there will be many different auxiliary information and fusion techniques that can be explored for few-shot and zero-shot information extraction. In this section, some research directions are discussed as follows:

### 7.3.1 Other Information Extraction Tasks

In this dissertation, we focus on a general relation extraction task including relation classification and relation triplet extraction. Besides, we explore aspect (attribute-value) extraction in

the e-commerce application field. Besides these information extraction tasks, there are some other tasks, such as event extraction, feedback/sentiment extraction from reviews, opinion extraction, and so on, that can be explored under the few-shot and zero-shot settings. In addition, relation extraction can be enriched with more information including adding timestamps for dynamic relation extraction, adding higher-order relationships for higher-order relation extraction, etc. Furthermore, besides the application field of e-commerce, structured information is also significant in other application fields such as finance, social media, biomedical and healthcare, etc.

### 7.3.2 Different Auxiliary Information

In this dissertation, we use auxiliary information from label descriptions, side information, images, knowledge graphs, user behavior data, product inventory data, optical characters, and so on. More auxiliary information can be explored to enhance few-shot and zero-shot information extraction tasks. For example, using data from external datasets that can provide supplementary information, combining data from multiple modalities such as images, text, audio, or sensor data that can provide a richer representation, implementing data augmentation such as including domain knowledge from expert feedback or interactions, and so on. More auxiliary information can be automatically collected by the web, generated by large language models, or provided by domain experts.

### 7.3.3 Zero-Shot Learning Exploration

In this dissertation, we implement zero-shot learning by leveraging various auxiliary information to support the missing semantics of the unlabeled data. In addition, we use data augmentation, semi-inductive link prediction, and large language models to perform zero-shot prediction. In future work, fine-tuning large language models with high-quality auxiliary data will be studied to generate zero-shot structured information. Besides, taking advantage of the encoder-decoder architecture of transformers, different decoding strategies can be explored in future work. For example, sampling-based polishing, Gibbs polishing, dynamic prompts, and other decoding strategies can be explored to enhance the zero-shot abilities for large language models while keeping effective and efficient of the fine-tuning process for large language models.

# Bibliography


[1] Azad Abad, Moin Nabi, and Alessandro Moschitti. Self-crowdsourcing training for relation extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 518–523, 2017.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[3] Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Laso: Label-set operations networks for multi-label few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6548–6557, 2019.

[4] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd ACL and the 7th IJCNL*, pages 344–354, July 2015.

[5] Douglas E Appelt and Boyan Onyshkevych. The common pattern specification language. Technical report, SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER, 1998.

[6] Douglas E Appelt, Jerry R Hobbs, John Bear, David Israel, and Mabry Tyson. Fastus: A finite-state processor for information extraction from real-world text. In *IJCAI*, volume 93, pages 1172–1178, 1993.

[7] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1279. URL https://aclanthology.org/P19-1279.

[8] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15338–15347, 2023.

[9] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018.

[10] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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 H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[12] Matteo Bustreo, Jacopo Cavazza, and Vittorio Murino. Enhancing visual embeddings through weakly supervised captioning for zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Oct 2019.

[13] Alberto Cetoli. Exploring the zero-shot limit of FewRel. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1447–1451, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.124. URL https://aclanthology.org/2020.coling-main.124.

[14] Binghui Chen and Weihong Deng. Hybrid-attention based decoupled metric learning for zero-shot image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2750–2759, 2019.

[15] Chih-Yao Chen and Cheng-Te Li. ZS-BERT: Towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3470–3479, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.272. URL https://aclanthology.org/2021.naacl-main.272.

[16] Chih-Yao Chen and Cheng-Te Li. Zs-bert: Towards zero-shot relation extraction with attribute representation learning. In *Proceedings of 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2021)*, 2021.

[17] 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-21*, pages 4366–4373. International Joint Conferences on

Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/597. URL https://doi.org/10.24963/ijcai.2021/597. Survey Track.

[18] Jiayi Chen and Aidong Zhang. Hgmf: heterogeneous graph-based fusion for multimodal data with incompleteness. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1295–1305, 2020.

[19] Mengru Chen, Chao Huang, Lianghao Xia, Wei Wei, Yong Xu, and Ronghua Luo. Heterogeneous graph contrastive learning for recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 544–552, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394079. doi: 10.1145/3539597.3570484. URL https://doi.org/10.1145/3539597.3570484.

[20] Qi Chen, Wei Wang, Kaizhu Huang, and Frans Coenen. Zero-shot text classification via knowledge graph embedding for social media data. *IEEE Internet of Things Journal*, pages 1–1, 2021. doi: 10.1109/JIOT.2021.3093065.

[21] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, Vancouver, July 2017. ACL.

[22] Wei-Te Chen, Yandi Xia, and Keiji Shinzato. Extreme multi-label classification with label masking for product attribute value extraction. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 134–140, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.ecnlp-1.16. URL https://aclanthology.org/2022.ecnlp-1.16.

[23] Wei-Te Chen, Yandi Xia, and Keiji Shinzato. Extreme multi-label classification with label masking for product attribute value extraction. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 134–140, 2022.

[24] Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. *CoRR*, abs/2104.07650, 2021. URL https://arxiv.org/abs/2104.07650.

[25] Xianyu Chen, Ming Jiang, and Qi Zhao. Self-distillation for few-shot image captioning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 545–555, 2021.

[26] Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z Pan, Zonggang Yuan, and Huajun Chen. Zero-shot visual question answering using knowledge graph. In *International Semantic Web Conference*, pages 146–162. Springer, 2021.

[27] Ziyi Chen, Yutong Gao, Congyan Lang, Lili Wei, Yidong Li, Hongzhe Liu, and Fayao Liu. Integrating topology beyond descriptions for zero-shot learning. *Pattern Recognition*, page 109738, 2023.

[28] Dian Cheng, Jiawei Chen, Wenjun Peng, Wenqin Ye, Fuyu Lv, Tao Zhuang, Xiaoyi Zeng, and Xiangnan He. Ihgnn: Interactive hypergraph neural network for personalized product search. In *Proceedings of the ACM Web Conference 2022*, pages 256–265, 2022.

[29] Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 45–57, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.5. URL https://aclanthology.org/2022.findings-acl.5.

[30] Hejie Cui, Rongmei Lin, Nasser Zalmout, Chenwei Zhang, Jingbo Shang, Carl Yang, and Xian Li. Pv2tea: Patching visual modality to textual-established information extraction, 2023.

[31] Lei Cui, Furu Wei, and Ming Zhou. Neural open information extraction. In *Proceedings of the 56th Association for Computational Linguistics*, pages 407–413, July 2018.

[32] Joe Davison, Joshua Feldman, and Alexander Rush. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1109. URL https://aclanthology.org/D19-1109.

[33] Zhongfen Deng, Wei-Te Chen, Lei Chen, and S Yu Philip. Ae-smnsmlc: Multi-label classification with semantic matching and negative label sampling for product attribute value extraction. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1816–1821. IEEE, 2022.

[34] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[36] Kaize Ding, Jianling Wang, Jundong Li, Dingcheng Li, and Huan Liu. Be more with less: Hypergraph attention networks for inductive text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4927–4936, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.399. URL https://aclanthology.org/2020.emnlp-main.399.

[37] Bowen Dong, Yuan Yao, Ruobing Xie, Tianyu Gao, Xu Han, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. Meta-information guided meta-learning for few-shot relation classification. In *Pro. of the 28th ICCL*, pages 1594–1605, 2020.

[38] Manqing Dong, Chunguang Pan, and Zhipeng Luo. MapRE: An effective semantic mapping approach for low-resource relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2694–2704, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.212. URL https://aclanthology.org/2021.emnlp-main.212.

[39] Xuanyi Dong, Linchao Zhu, De Zhang, Yi Yang, and Fei Wu. Fast parameter adaptation for few-shot image captioning and visual question answering. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 54–62, 2018.

[40] Cícero dos Santos, Bing Xiang, and Bowen Zhou. Classifying relations by ranking with convolutional neural networks. In *Proc. of the 53rd Association for Computational Linguistics and the 7th IJCNLP*, pages 626–634, July 2015.

[41] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[42] Chunliu Dou, Shaojuan Wu, Xiaowang Zhang, Zhiyong Feng, and Kewen Wang. Function-words adaptively enhanced attention networks for few-shot inverse relation classification. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2937–2943. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/407. URL https://doi.org/10.24963/ijcai.2022/407. Main Track.

[43] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual*

*SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5506. URL https://aclanthology.org/W17-5506.

[44] Haoyi Fan, Fengbin Zhang, Yuxuan Wei, Zuoyong Li, Changqing Zou, Yue Gao, and Qionghai Dai. Heterogeneous hypergraph variational autoencoder for link prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4125–4138, 2021.

[45] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2020.

[46] Ziwei Fan, Zhiwei Liu, Shelby Heinecke, Jianguo Zhang, Huan Wang, Caiming Xiong, and Philip S Yu. Zero-shot item-based recommendation via multi-task product knowledge graph pre-training. *arXiv preprint arXiv:2305.07633*, 2023.

[47] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3136–3146, 2023.

[48] Yu Fei, Zhao Meng, Ping Nie, Roger Wattenhofer, and Mrinmaya Sachan. Beyond prompting: Making pre-trained language models better zero-shot learners by clustering representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8560–8579, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.587.

[49] Dayne Freitag, John Cadigan, Robert Sasseen, and Paul Kalmar. Valet: Rule-based information extraction for rapid deployment. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 524–533, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.55.

[50] Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. Hybrid attention-based prototypical networks for noisy few-shot relation classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6407–6414, Jul. 2019. doi: 10.1609/aaai.v33i01.33016407. URL https://ojs.aaai.org/index.php/AAAI/article/view/4604.

[51] Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. Hybrid attention-based prototypical networks for noisy few-shot relation classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6407–6414, 07 2019.

[52] Tianyu Gao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. Neural snowball for few-shot relation learning. In *AAAI*, 2020.

[53] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Association for Computational Linguistics (ACL)*, 2021.

[54] Yue Gao, Yifan Feng, Shuyi Ji, and Rongrong Ji. Hgnn+: General hypergraph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3181–3199, 2023. doi: 10.1109/TPAMI.2022.3182052.

[55] Zhi Gao, Yuwei Wu, Yunde Jia, and Mehrtash Harandi. Curvature generation in curved spaces for few-shot learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8671–8680, 2021. doi: 10.1109/ICCV48922.2021.00857.

[56] Xiaoqing Geng, Xiwen Chen, Kenny Q. Zhu, Libin Shen, and Yinggong Zhao. Mick: A meta-learning framework for few-shot relation classification with small training data. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 415–424, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3411858. URL https://doi.org/10.1145/3340531.3411858.

[57] Yuxia Geng, Jiaoyan Chen, Wen Zhang, Yajing Xu, Zhuo Chen, Jeff Z. Pan, Yufeng Huang, Feiyu Xiong, and Huajun Chen. Disentangled ontology embedding for zero-shot learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 443–453, 2022.

[58] Yuxia Geng, Jiaoyan Chen, Xiang Zhuang, Zhuo Chen, Jeff Z Pan, Juan Li, Zonggang Yuan, and Huajun Chen. Benchmarking knowledge-driven zero-shot learning. *Journal of Web Semantics*, 75:100757, 2023.

[59] Rayid Ghani, Katharina Probst, Yan Liu, Marko Krema, and Andrew Fano. Text mining for product attribute extraction. *SIGKDD Explor. Newsl.*, 8(1):41–48, jun 2006. ISSN 1931-0145. doi: 10.1145/1147234.1147241. URL https://doi.org/10.1145/1147234.1147241.

[60] Pushpendu Ghosh, Nancy Wang, and Promod Yenigalla. D-extract: Extracting dimensional attributes from product images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3641–3649, 2023.

[61] Jiaying Gong and Hoda Eldardiry. *Zero-Shot Relation Classification from Side Information*, page 576–585. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450384469. URL https://doi.org/10.1145/3459637.3482403.

[62] Jiaying Gong and Hoda Eldardiry. Prompt-based zero-shot relation extraction with semantic knowledge augmentation, 2023.

[63] Jiaying Gong, Wei-Te Chen, and Hoda Eldardiry. Knowledge-enhanced multi-label few-shot product attribute-value extraction. *arXiv preprint arXiv:2308.08413*, 2023.

[64] Peizhu Gong, Jin Liu, Xiliang Zhang, Xingye Li, and Zijun Yu. Circulant-interactive transformer with dimension-aware fusion for multimodal sentiment analysis. In *Asian Conference on Machine Learning*, pages 391–406. PMLR, 2023.

[65] Dagmar Gromann, Yuxia Geng, Jiaoyan Chen, Zhiquan Ye, Zonggang Yuan, Wei Zhang, and Huajun Chen. Explainable zero-shot learning via attentive graph convolutional network and knowledge graphs. *Semant. Web*, 12(5):741–765, jan 2021. ISSN 1570-0844. doi: 10.3233/SW-210435. URL https://doi.org/10.3233/SW-210435.

[66] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. Ppt: Pre-trained prompt tuning for few-shot learning, 2021.

[67] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877, 2023.

[68] Sheng Guo, Weilin Huang, Xiao Zhang, Prasanna Srikhanta, Yin Cui, Yuan Li, Matthew R. Scott, Hartwig Adam, and Serge J. Belongie. The imaterialist fashion attribute dataset. *CoRR*, abs/1906.05750, 2019. URL http://arxiv.org/abs/1906.05750.

[69] Pietro Hiram Guzzi and Marinka Zitnik. Editorial deep learning and graph embeddings for network biology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(2):653–654, 2022.

[70] Jiale Han, Bo Cheng, and Wei Lu. Exploring task difficulty for few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2605–2616, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.204. URL https://aclanthology.org/2021.emnlp-main.204.

[71] Jiale Han, Bo Cheng, and Wei Lu. Exploring task difficulty for few-shot relation extraction. In *Proc. of EMNLP*, 2021.

[72] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, October-November 2018.

[73] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. Ptr: Prompt tuning with rules for text classification, 2021.

[74] Zhongxuan Han, Xiaolin Zheng, Chaochao Chen, Wenjie Cheng, and Yang Yao. Intra and inter domain hypergraph convolutional network for cross-domain recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 449–459, 2023.

[75] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[76] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proc. of the 5th Inter. Workshop on Semantic Evaluation*, July 2010.

[77] Yutai Hou, Yongkui Lai, Yushan Wu, Wanxiang Che, and Ting Liu. Few-shot learning for multi-label intent detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13036–13044, 2021.

[78] Linmei Hu, Luhao Zhang, Chuan Shi, Liqiang Nie, Weili Guan, and Cheng Yang. Improving distantly-supervised relation extraction with joint label embedding. In *Proc. of Conference on EMNLP and the 9th IJCNL*, pages 3821–3829, November 2019.

[79] Mengting Hu, Shiwan Zhao, Honglei Guo, Chao Xue, Hang Gao, Tiegang Gao, Renhong Cheng, and Zhong Su. Multi-label few-shot learning for aspect category detection. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6330–6340. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.495. URL https://doi.org/10.18653/v1/2021.acl-long.495.

[80] Mengting Hu, Shiwan Zhao, Honglei Guo, Chao Xue, Hang Gao, Tiegang Gao, Renhong Cheng, and Zhong Su. Multi-label few-shot learning for aspect category detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6330–6340, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.495. URL https://aclanthology.org/2021.acl-long.495.

[81] Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Juanzi Li, and Maosong Sun. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification, 2021.

[82] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, pages 2704–2710, 2020.

[83] Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. Zero-shot transfer learning for event extraction. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2160–2170, July 2018.

[84] Yen-Hao Huang, Yi-Hsin Chen, and Yi-Shin Chen. Contexting: Granting document-wise contextual embeddings to graph neural networks for inductive text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1163–1168, 2022.

[85] Yucheng Huang, Kai He, Yige Wang, Xianli Zhang, Tieliang Gong, Rui Mao, and Chen Li. COPNER: Contrastive learning with prompt guiding for few-shot named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2515–2527, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.222.

[86] Pere-Lluís Huguet Cabot and Roberto Navigli. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.204. URL https://aclanthology.org/2021.findings-emnlp.204.

[87] Bei Hui, Liang Liu, Jia Chen, Xue Zhou, and Yuhui Nian. Few-shot relation classification by context attention-based prototypical networks with bert. *EURASIP Journal on Wireless Communications and Networking*, 2020, 12 2020.

[88] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE/CVF*, June 2020.

[89] Mayank Jain, Sourangshu Bhattacharya, Harshit Jain, Karimulla Shaik, and Muthusamy Chelliah. Learning cross-task attribute - attribute similarity for multi-task attribute-value extraction. In *Proceedings of the 4th Workshop on e-Commerce and NLP*, pages 79–87, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.ecnlp-1.10. URL https://aclanthology.org/2021.ecnlp-1.10.

[90] Shengbin Jia and Yang Xiang. Hybrid neural tagging model for open relation extraction. *arXiv: Computation and Language*, 2020.

[91] Guangyuan Jiang, Manjie Xu, Shiji Xin, Wei Liang, Yujia Peng, Chi Zhang, and Yixin Zhu. Mewl: Few-shot multimodal word learning with referential uncertainty. *arXiv preprint arXiv:2306.00503*, 2023.

[92] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438, 2020. URL https://transacl.org/ojs/index.php/tacl/article/view/1983.

[93] Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. Gaze embeddings for zero-shot image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4525–4534, 2017.

[94] Rohit Keshari, Richa Singh, and Mayank Vatsa. Generalized zero-shot learning via over-complete distribution. In *Proceedings of the IEEE/CVF*, June 2020.

[95] Bilal Khan, Jia Wu, Jian Yang, and Xiaoxiao Ma. Heterogeneous hypergraph neural network for social recommendation using attention network. *ACM Transactions on Recommender Systems*, 2023.

[96] Bosung Kim, Hayate Iso, Nikita Bhutani, Estevam Hruschka, and Ndapa Nakashole. Zero-shot triplet extraction by template infilling. *arXiv preprint arXiv:2212.10708*, 2022.

[97] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[98] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille, 2015.

[99] Yuquan Lan, Dongxu Li, Hui Zhao, and Gang Zhao. Pcred: Zero-shot relation triplet extraction with potential candidate relation selection and entity boundary detection. *arXiv preprint arXiv:2211.14477*, 2022.

[100] Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-1618. URL https://aclanthology.org/W14-1618.

[101] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pages 333–342, August 2017.

[102] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[103] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. In *Proc. of the IEEE/CVF*, June 2020.

[104] Cheng Li and Ye Tian. Downstream model design of pre-trained language model for relation extraction task. *CoRR*, abs/2004.03786, 2020. URL https://arxiv.org/abs/2004.03786.

[105] Cheng Li and Ye Tian. Downstream model design of pre-trained language model for relation extraction task. *ArXiv*, 2020.

[106] Dongfang Li, Baotian Hu, and Qingcai Chen. Prompt-based text entailment for low-resource named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1896–1903, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.164.

[107] Haochen Li, Tong Mo, Hongcheng Fan, Jingkun Wang, Jiaxi Wang, Fuhao Zhang, and Weiping Li. KiPT: Knowledge-injected prompt tuning for event detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1943–1952, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.169.

[108] Juan Li, Ruoxu Wang, Ningyu Zhang, Wen Zhang, Fan Yang, and Huajun Chen. Logic-guided semantic representation learning for zero-shot relation classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2967–2978, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.265. URL https://aclanthology.org/2020.coling-main.265.

[109] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[110] Kai Li, Martin Renqiang Min, and Yun Fu. Rethinking zero-shot learning: A conditional visual classification perspective. *2019 IEEE/CVF*, pages 3582–3591, 2019.

[111] Mengran Li, Yong Zhang, Xiaoyong Li, Yuchen Zhang, and Baocai Yin. Hypergraph transformer neural networks. *ACM Transactions on Knowledge Discovery from Data*, 17(5):1–22, 2023.

[112] Wanli Li and Tieyun Qian. Graph-based model generation for few-shot relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 62–71, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.5.

[113] Yongkang Li, Zipei Fan, Du Yin, Renhe Jiang, Jinliang Deng, and Xuan Song. Hmgcl: Heterogeneous multigraph contrastive learning for lbsn friend recommendation. *World Wide Web*, pages 1–24, 2022.

[114] Yongkang Li, Zipei Fan, Jixiao Zhang, Dengheng Shi, Tianqi Xu, Du Yin, Jinliang Deng, and Xuan Song. Heterogeneous hypergraph neural network for friend recommendation with human mobility. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4209–4213, 2022.

[115] Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, and Andrei Bursuc. Dense classification and implanting for few-shot learning. In *Proceedings of the IEEE/CVF*, June 2019.

[116] Rongmei Lin, Xiang He, Jie Feng, Nasser Zalmout, Yan Liang, Li Xiong, and Xin Luna Dong. Pam: understanding product images in cross product category attribute extraction. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3262–3270, 2021.

[117] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

[118] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19325–19337, 2023.

[119] Xiao Ling and Daniel S. Weld. Fine-grained entity recognition. In *Proc. of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, page 94–100. AAAI Press, 2012.

[120] Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H Lin, Xiao Ling, and Daniel S Weld. Effective crowd annotation for relation extraction. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 897–906, 2016.

[121] ChunYang Liu, WenBo Sun, WenHan Chao, and WanXiang Che. Convolution neural network for relation extraction. pages 231–242, 12 2013.

[122] Fangchao Liu, Hongyu Lin, Xianpei Han, Boxi Cao, and Le Sun. Pre-training to match for unified low-shot relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5785–5795, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.397. URL https://aclanthology.org/2022.acl-long.397.

[123] Han Liu, Feng Zhang, Xiaotong Zhang, Siyang Zhao, Junjie Sun, Hong Yu, and Xianchao Zhang. Label-enhanced prototypical network with contrastive learning for multi-label few-shot aspect category detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1079–1087, 2022.

[124] Jie Liu, Lingyun Song, Guangtao Wang, and Xuequn Shang. Meta-hgt: Metapath-aware hypergraph transformer for heterogeneous information network embedding. *Neural Networks*, 157:65–76, 2023.

[125] Jingquan Liu, Xiaoyong Du, Yuanzhe Li, and Weidong Hu. Hypergraph variational autoencoder for multimodal semi-supervised representation learning. In *Artificial Neural Networks and Machine Learning–ICANN 2022: 31st International Conference on Artificial Neural Networks, Bristol, UK, September 6–9, 2022, Proceedings; Part IV*, pages 395–406. Springer, 2022.

[126] Jiongnan Liu, Zhicheng Dou, Guoyu Tang, and Sulong Xu. Jdsearch: A personalized product search dataset with real queries and full interactions. In *Proceedings of the SIGIR 2023*. ACM, 2023. doi: 10.1145/3539618.3591900. URL https://doi.org/10.1145/3539618.3591900.

[127] Mengyin Liu, Chao Zhu, Hongyu Gao, Weibo Gu, Hongfa Wang, Wei Liu, and Xu cheng Yin. Boosting multi-modal e-commerce attribute value extraction via unified learning scheme and dynamic range minimization, 2023.

[128] Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. Hyperbolic visual embedding learning for zero-shot recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[129] Shilei Liu, Lin Li, Jun Song, Yonghua Yang, and Xiaoyi Zeng. Multimodal pre-training with self-distillation for product understanding in e-commerce. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1039–1047, 2023.

[130] Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. Learn from relation information: Towards prototype representation rectification for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1822–1831, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.139. URL https://aclanthology.org/2022.findings-naacl.139.

[131] Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. A simple yet effective relation information guided approach for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 757–763, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.62. URL https://aclanthology.org/2022.findings-acl.62.

[132] Yun Liu, Xiaoming Zhang, Qianyun Zhang, Chaozhuo Li, Feiran Huang, Xianghong Tang, and Zhoujun Li. Dual self-attention with co-attention networks for visual question answering. *Pattern Recognition*, 117:107956, 2021.

[133] Colin Lockard, Prashant Shiralkar, Xin Luna Dong, and Hannaneh Hajishirzi. ZeroShotCeres: Zero-shot relation extraction from semi-structured webpages. In *Pro. of the 58th Association for Computational Linguistics*, pages 8105–8117, July 2020.

[134] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1185. URL https://aclanthology.org/P18-1185.

[135] Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Qi Zhang, and Xuanjing Huang. Template-free prompt tuning for few-shot ner, 2021.

[136] Youmi Ma, Tatsuya Hiraoka, and Naoaki Okazaki. Joint entity and relation extraction based on table labeling using convolutional neural networks. In *Proceedings of the Sixth Workshop on Structured Prediction for NLP*, pages 11–21, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.spnlp-1.2. URL https://aclanthology.org/2022.spnlp-1.2.

[137] Yuqing Ma, Shihao Bai, Shan An, Wei Liu, Aishan Liu, Xiantong Zhen, and Xianglong Liu. Transductive relation-propagation network for few-shot learning. In *IJCAI*, volume 20, pages 804–810, 2020.

[138] Naveen Madapana. Zero-shot learning for gesture recognition. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 754–757, 2020.

[139] Diego Marcheggiani and Ivan Titov. Discrete-state variational autoencoders for joint discovery and factorization of relations. *Transactions of the Association for Computational Linguistics*, 4:231–244, 2016.

[140] Eirinaios Michelakis, Rajasekar Krishnamurthy, Peter J Haas, and Shivakumar Vaithyanathan. Uncertainty management in rule-based information extraction systems. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 101–114, 2009.

[141] Marianna Milano, Giuseppe Agapito, and Mario Cannataro. Challenges and limitations of biological network analysis. *BioTech*, 11(3):24, 2022.

[142] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, 3(4):235–244, 12 1990.

[143] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009.

[144] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[145] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1DmUzWAW.

[146] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*, 2023.

[147] Muhammad Ferjad Naeem, Yongqin Xian, Luc V Gool, and Federico Tombari. I2dformer: Learning image to document attention for zero-shot image classification. *Advances in Neural Information Processing Systems*, 35:12283–12294, 2022.

[148] Ivona Najdenkoska, Xiantong Zhen, and Marcel Worring. Meta learning to bridge vision and language models for multimodal few-shot learning. *arXiv preprint arXiv:2302.14794*, 2023.

[149] Thien Huu Nguyen and Ralph Grishman. Combining neural networks and log-linear models to improve relation extraction. *CoRR*, abs/1511.05926, 2015.

[150] Thien Huu Nguyen and Ralph Grishman. Relation extraction: Perspective from convolutional neural networks. In *Proc. of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, June 2015.

[151] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197, 2019.

[152] Xinzhe Ni, Hao Wen, Yong Liu, Yatai Ji, and Yujiu Yang. Multimodal prototype-enhanced network for few-shot action recognition. *arXiv preprint arXiv:2212.04873*, 2022.

[153] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *ArXiv*, 2018.

[154] Abiola Obamuyide and Andreas Vlachos. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5511. URL https://aclanthology.org/W18-5511.

[155] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[156] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250.

[157] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and Q. M. Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4051–4070, 2023. doi: 10.1109/TPAMI.2022.3191696.

[158] Duangmanee (Pew) Putthividhya and Junling Hu. Bootstrapped named entity recognition for product attribute extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 1557–1567, USA, 2011. Association for Computational Linguistics. ISBN 9781937284114.

[159] Chengwei Qin and Shafiq Joty. Continual few-shot relation learning via embedding space regularization and data augmentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2776–2789, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.198. URL https://aclanthology.org/2022.acl-long.198.

[160] Pengda Qin, Xin Wang, Wenhu Chen, Chunyun Zhang, Weiran Xu, and William Yang Wang. Generative adversarial zero-shot relational learning for knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8673–8680, Apr. 2020. doi: 10.1609/aaai.v34i05.6392. URL https://ojs.aaai.org/index.php/AAAI/article/view/6392.

[161] Meng Qu, Tianyu Gao, Louis-Pascal AC Xhonneux, and Jian Tang. Few-shot relation extraction via bayesian meta-learning on relation graphs. In *International Conference on Machine Learning*, 2020.

[162] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[163] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[164] Rahul Ragesh, Sundararajan Sellamanickam, Arun Iyer, Ramakrishna Bairi, and Vijay Lingam. Hetegcn: heterogeneous graph convolutional networks for text classification. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 860–868, 2021.

[165] Elahe Rahimian, Soheil Zabihi, Amir Asif, S Farokh Atashzar, and Arash Mohammadi. Few-shot learning for decoding surface electromyography for hand gesture recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1300–1304. IEEE, 2021.

[166] Shafin Rahman, Salman Khan, and Nick Barnes. Transductive learning for zero-shot object detection. In *Proceedings of the IEEE/CVF*, October 2019.

[167] Juan Ramos. Using tf-idf to determine word relevance in document queries. 01 2003.

[168] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. Multi-modal zero-shot dynamic hand gesture recognition. *Expert Systems with Applications*, 247:123349, 2024.

[169] Frederick Reiss, Sriram Raghavan, Rajasekar Krishnamurthy, Huaiyu Zhu, and Shivakumar Vaithyanathan. An algebraic approach to rule-based information extraction. In *2008 IEEE 24th International Conference on Data Engineering*, pages 933–942. IEEE, 2008.

[170] Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. A novel global feature-oriented relational triple extraction model based on table filling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2646–2656, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.208. URL https://aclanthology.org/2021.emnlp-main.208.

[171] Haopeng Ren, Yi Cai, Xiaofeng Chen, Guohua Wang, and Qing Li. A two-phase prototypical network model for incremental few-shot relation classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1618–1629, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.142. URL https://aclanthology.org/2020.coling-main.142.

[172] Martin Rezk, Laura Alonso Alemany, Lasguido Nio, and Ted Zhang. Accurate product attribute extraction on the field. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1862–1873, 2019. doi: 10.1109/ICDE.2019.00202.

[173] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In José Luis Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *MLKDD*, pages 148–163, 2010.

[174] Anthony Rios and Ramakanth Kavuluru. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, page 3132. NIH Public Access, 2018.

[175] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. Reasoning about entailment with neural attention. 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. URL http://arxiv.org/abs/1509.06664.

[176] Mohammad Rostami, Soheil Kolouri, Zak Murez, Yuri Owechko, Eric Eaton, and Kuyngnam Kim. Zero-shot image classification using coupled dictionary embedding. *Machine Learning with Applications*, 8:100278, 2022.

[177] Kalyani Roy, Pawan Goyal, and Manish Pandey. Attribute value generation from product title using language models. In *Proceedings of The 4th Workshop on e-Commerce and NLP*, pages 13–17, 2021.

[178] Kalyani Roy, Tapas Nayak, and Pawan Goyal. Exploring generative models for joint attribute value extraction from product titles. *arXiv preprint arXiv:2208.07130*, 2022.

[179] Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.92. URL https://aclanthology.org/2021.emnlp-main.92.

[180] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJj6qGbRW.

[181] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[182] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 255–269. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.eacl-main.20/.

[183] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2339–2352. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.185. URL https://doi.org/10.18653/v1/2021.naacl-main.185.

[184] Xi Shen, Yang Xiao, Shell Xu Hu, Othman Sbai, and Mathieu Aubry. Re-ranking for image retrieval and transductive few-shot classification. *Advances in Neural Information Processing Systems*, 34:25932–25943, 2021.

[185] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. 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 4222–4235. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.346. URL https://doi.org/10.18653/v1/2020.emnlp-main.346.

[186] Keiji Shinzato and Satoshi Sekine. Unsupervised extraction of attributes and their values from product description. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1339–1347, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I13-1190.

[187] Keiji Shinzato, Naoki Yoshinaga, Yandi Xia, and Wei-Te Chen. Simple and effective knowledge-driven query expansion for qa-based product attribute extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 227–234, 2022.

[188] Keiji Shinzato, Naoki Yoshinaga, Yandi Xia, and Wei-Te Chen. A unified generative approach to product attribute-value identification. *arXiv preprint arXiv:2306.05605*, 2023.

[189] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. Meta-learning for multi-label few-shot classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3951–3960, 2022.

[190] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS 30*, pages 4077–4087. 2017.

[191] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf.

[192] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[193] Daniil Sorokin and Iryna Gurevych. Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1784–1789, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1188. URL https://aclanthology.org/D17-1188.

[194] Robyn Speer and Catherine Havasi. *ConceptNet 5: A Large Semantic Network for Relational Knowledge*, pages 161–176. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-35085-6. doi: 10.1007/978-3-642-35085-6_6. URL https://doi.org/10.1007/978-3-642-35085-6_6.

[195] Lin Sun, Jiquan Wang, Yindu Su, Fangsheng Weng, Yuxuan Sun, Zengwei Zheng, and Yuanyi Chen. RIVA: A pre-trained tweet multimodal model based on text-image relation for multimodal NER. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1852–1862, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.168. URL https://aclanthology.org/2020.coling-main.168.

[196] Xiangguo Sun, Hongzhi Yin, Bo Liu, Hongxu Chen, Jiuxin Cao, Yingxia Shao, and Nguyen Quoc Viet Hung. Heterogeneous hypergraph embedding for graph classification. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 725–733, 2021.

[197] 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 *European Conference on Computer Vision*, pages 266–282. Springer, 2020.

[198] Trieu H. Trinh and Quoc V. Le. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847, 2018. URL http://arxiv.org/abs/1806.02847.

[199] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[200] Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. RESIDE: Improving distantly-supervised neural relation extraction

using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266, October-November 2018.

[201] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[202] Hai Wan, Manrong Zhang, Jianfeng Du, Ziling Huang, Yufei Yang, and Jeff Z Pan. Fl-msre: A few-shot learning based approach to multimodal social relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13916–13923, 2021.

[203] Jue Wang and Wei Lu. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.133. URL https://aclanthology.org/2020.emnlp-main.133.

[204] Kai Wang, Jianzhi Shao, Tao Zhang, Qijin Chen, and Chengfu Huo. Mpkgac: Multimodal product attribute completion in e-commerce. In *Companion Proceedings of the ACM Web Conference 2023*, pages 336–340, 2023.

[205] Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1298–1307, August 2016.

[206] Qifan Wang, Li Yang, Bhargav Kanagal, Sumit Sanghai, D Sivakumar, Bin Shu, Zac Yu, and Jon Elsas. Learning to extract attribute value from product via question answering: A multi-task approach. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 47–55, 2020.

[207] Qifan Wang, Li Yang, Jingang Wang, Jitin Krishnan, Bo Dai, Sinong Wang, Zenglin Xu, Madian Khabsa, and Hao Ma. Smartave: Structured multimodal transformer for product attribute value extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 263–276, 2022.

[208] Shusen Wang, Bosen Zhang, Yajing Xu, Yanan Wu, and Bo Xiao. RCL: Relation contrastive learning for zero-shot relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2456–2468, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.188. URL https://aclanthology.org/2022.findings-naacl.188.

[209] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Heterogeneous graph attention network. In *The world wide web conference*, pages 2022–2032, 2019.

[210] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, 53(3), jun 2020. ISSN 0360-0300. doi: 10.1145/3386252. URL https://doi.org/10.1145/3386252.

[211] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3): 1–34, 2020.

[212] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *Advances in neural information processing systems*, 33:4835–4845, 2020.

[213] Yingyao Wang, Junwei Bao, Guangyi Liu, Youzheng Wu, Xiaodong He, Bowen Zhou, and Tiejun Zhao. Learning to decouple relations: Few-shot relation classification with entity-guided attention and confusion-aware training. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5799–5809. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.510. URL https://doi.org/10.18653/v1/2020.coling-main.510.

[214] Yuyang Wanyan, Xiaoshan Yang, Chaofan Chen, and Changsheng Xu. Active exploration of multimodal complementarity for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6492–6502, 2023.

[215] Xuemei Wei, Yezheng Liu, Jianshan Sun, Yuanchun Jiang, Qifeng Tang, and Kun Yuan. Dual subgraph-based graph neural network for friendship prediction in location-based social networks. *ACM Transactions on Knowledge Discovery from Data*, 17(3): 1–28, 2023.

[216] Jamaal Hay Wenpeng Yin and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *EMNLP*, 2019. URL https://arxiv.org/abs/1909.00161.

[217] Zheng Wenping, Liu Meilin, and Liang Jiye. Hypergraphs: Concepts, applications and analysis. In *2022 IEEE 13th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP)*, pages 1–6, 2022. doi: 10.1109/PAAP56126.2022.10010428.

[218] J. D. Williams, A. Raux, D. Ramachandran, and A. Black. Dialog state tracking challenge handbook. In *Technical report, Technical report, Microsoft Research.*, 2012.

[219] Yuk Wah Wong, Dominic Widdows, Tom Lokovic, and Kamal Nigam. Scalable attribute-value extraction from semi-structured text. In *ICDM Workshop on Large-scale Data Mining: Theory and Applications*, 2009. URL http://www.computer.org/portal/web/csdl/doi/10.1109/ICDMW.2009.81.

[220] Aming Wu and Yahong Han. Multi-modal circulant fusion for video-to-language and backward. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1029–1035. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/143. URL https://doi.org/10.24963/ijcai.2018/143.

[221] Hanrui Wu, Yuguang Yan, and Michael Kwok-Po Ng. Hypergraph collaborative network on vertices and hyperedges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3245–3258, 2022.

[222] Ruidong Wu, Yuan Yao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. Open relation extraction: Relational knowledge transfer from supervised data to unsupervised data. In *Proceedings of the 2019 Conference on EMNLP and the 9th IJCNP*, pages 219–228, November 2019.

[223] Shanchan Wu and Yifan He. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2361–2364, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369763. doi: 10.1145/3357384.3358119. URL https://doi.org/10.1145/3357384.3358119.

[224] Haoke Xiao, Lv Tang, Bo Li, Zhiming Luo, and Shaozi Li. Zero-shot co-salient object detection framework. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4010–4014. IEEE, 2024.

[225] Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao. Attentive region embedding network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019.

[226] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. *arXiv preprint arXiv:2002.07962*, 2020.

[227] Huimin Xu, Wenting Wang, Xinnian Mao, Xinyu Jiang, and Man Lan. Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5214–5223, 2019.

[228] Liyan Xu, Chenwei Zhang, Xian Li, Jingbo Shang, and Jinho D Choi. Towards open-world product attribute mining: A lightly-supervised approach. *arXiv preprint arXiv:2305.18350*, 2023.

[229] Zixuan Xu, Penghui Wei, Shaoguo Liu, Weimin Zhang, Liang Wang, and Bo Zheng. Correlative preference transfer with hierarchical hypergraph network for multi-domain recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 983–991, 2023.

[230] Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. Hypergcn: A new method for training graph convolutional networks on hypergraphs. *Advances in neural information processing systems*, 32, 2019.

[231] Jun Yan, Nasser Zalmout, Yan Liang, Christan Grant, Xiang Ren, and Xin Luna Dong. AdaTag: Multi-attribute value extraction from product profiles with adaptive decoding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4694–4705, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.362. URL https://aclanthology.org/2021.acl-long.362.

[232] Kun Yan, Chenbin Zhang, Jun Hou, Ping Wang, Zied Bouraoui, Shoaib Jameel, and Steven Schockaert. Inferring prototypes for multi-label few-shot image classification with word vector guided attention. *36th AAAI Conference on Artificial Intelligence.*, 2022.

[233] Kun Yan, Chenbin Zhang, Jun Hou, Ping Wang, Zied Bouraoui, Shoaib Jameel, and Steven Schockaert. Inferring prototypes for multi-label few-shot image classification with word vector guided attention. 2022.

[234] Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. Mave: A product dataset for multi-source attribute value extraction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 1256–1265, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391320. doi: 10.1145/3488560.3498377. URL https://doi.org/10.1145/3488560.3498377.

[235] Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. Mave: A product dataset for multi-source attribute value extraction. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 1256–1265, 2022.

[236] Shan Yang, Yongfei Zhang, Guanglin Niu, Qinghua Zhao, and Shiliang Pu. Entity concept-enhanced few-shot relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 987–991, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.124. URL https://aclanthology.org/2021.acl-short.124.

[237] Xiaocui Yang, Shi Feng, Daling Wang, Pengfei Hong, and Soujanya Poria. Few-shot multimodal sentiment analysis based on multimodal probabilistic fusion prompts. *arXiv preprint arXiv:2211.06607*, 2022.

[238] Zhuo Yang, Yufei Han, Guoxian Yu, Qiang Yang, and Xiangliang Zhang. Prototypical networks for multi-label learning. *arXiv preprint arXiv:1911.07203*, 2019.

[239] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF*, June 2020.

[240] Zhi-Xiu Ye and Zhen-Hua Ling. Multi-level matching and aggregation network for few-shot relation classification. In *Proc. of the 57th ACL*, July 2019.

[241] Tan Yu, Yi Yang, Yi Li, Lin Liu, Hongliang Fei, and Ping Li. *Heterogeneous Attention Network for Effective and Efficient Cross-Modal Retrieval*, page 1146–1156. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450380379. URL https://doi.org/10.1145/3404835.3462924.

[242] Yunlong Yu, Zhong Ji, Jungong Han, and Zhongfei Zhang. Episode-based prototype generating network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

[243] J. Yuan, H. Guo, Z. Jin, H. Jin, X. Zhang, and J. Luo. One-shot learning for fine-grained relation extraction via convolutional siamese neural network. In *2017 IEEE International Conference on Big Data*, pages 2194–2199, 2017.

[244] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014*, pages 2335–2344, August 2014.

[245] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762, 2015.

[246] Dongxu Zhang and Dong Wang. Relation classification via recurrent neural network. *CoRR*, abs/1508.01006, 2015.

[247] Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. Integrating semantic knowledge to tackle zero-shot text classification. In *Proc. of NAACL*, June 2019.

[248] Jiying Zhang, Yuzhao Chen, Xi Xiao, Runiu Lu, and Shu-Tao Xia. Learnable hypergraph laplacian for hypergraph learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4503–4507. IEEE, 2022.

[249] Liyan Zhang, Jingfeng Guo, Jiazheng Wang, Jing Wang, Shanshan Li, and Chunying Zhang. Hypergraph and uncertain hypergraph representation learning theory and methods. *Mathematics*, 10(11):1921, 2022.

[250] Peiyuan Zhang and Wei Lu. Better few-shot relation extraction with label prompt dropout. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6996–7006, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.471.

[251] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. Adaptive co-attention network for named entity recognition in tweets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11962.

[252] Runyan Zhang, Fanrong Meng, Yong Zhou, and Bing Liu. Relation classification via recurrent neural network with attention and tensor layers. *Big Data Mining and Analytics*, 1(3):234–244, 2018.

[253] Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 73–78, Shanghai, China, October 2015. URL https://aclanthology.org/Y15-1009.

[254] Xinyang Zhang, Chenwei Zhang, Xian Li, Xin Luna Dong, Jingbo Shang, Christos Faloutsos, and Jiawei Han. Oa-mine: Open-world attribute mining for e-commerce products with weak supervision. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3153–3161, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512035. URL https://doi.org/10.1145/3485447.3512035.

[255] Yuhao Zhang, Peng Qi, and Christopher D. Manning. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on EMNLP*, pages 2205–2215, October-November 2018.

[256] Yupeng Zhang, Shensi Wang, Peiguang Li, Guanting Dong, Sirui Wang, Yunsen Xian, Zhoujun Li, and Hongzhi Zhang. Pay attention to implicit attribute values: a multi-modal generative framework for ave task. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13139–13151, 2023.

[257] Fei Zhao, Yuchen Shen, Zhen Wu, and Xinyu Dai. Label-driven denoising framework for multi-label few-shot aspect category detection. *arXiv preprint arXiv:2210.04220*, 2022.

[258] Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. *Multimodal Relation Extraction with Efficient Graph Alignment*, page 5298–5306. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450386517. URL https://doi.org/10.1145/3474085.3476968.

[259] Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. Mnre: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[260] Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. Opentag: Open attribute value extraction from product profiles. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, page 1049–1058, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219839. URL https://doi.org/10.1145/3219819.3219839.

[261] Li Zhenzhen, Yuyang Zhang, Jian-Yun Nie, and Dongsheng Li. Improving few-shot relation classification by prototypical representation learning with definition text. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 454–464, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.34. URL https://aclanthology.org/2022.findings-naacl.34.

[262] Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.5. URL https://aclanthology.org/2021.naacl-main.5.

[263] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th ACL*, pages 207–212, August 2016.

[264] Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. Graph neural networks with generated parameters for relation extraction. In *Proceedings of the 57th Association for Computational Linguistics*, pages 1331–1339, July 2019.

[265] Jizhao Zhu, Jianzhong Qiao, Xinxiao Dai, and Xueqi Cheng. Relation classification via target-concentrated attention cnns. pages 137–146, 10 2017.

[266] Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. Multimodal joint attribute prediction and value extraction for E-commerce product. In *Proceedings of the 2020 Conference on Empirical Methods in Natural*

*Language Processing (EMNLP)*, pages 2129–2139, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.166. URL https://aclanthology.org/2020.emnlp-main.166.