CROSS-DOMAIN FEW-SHOT RELATION EXTRACTION VIA REPRESENTATION LEARNING AND DOMAIN ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot relation extraction aims to recognize novel relations with few labeled sentences in each relation. Previous metric-based few-shot relation extraction methods classify by comparing the embeddings of query sentence embedding with those prototypes generated by the few labeled sentences embedding using a learned metric function. However, the generalization ability of these methods on unseen relations in different domains is limited, since these domains always have significant discrepancies from those in the training dataset. Because the prototype is essential for extracting relations between entities in the latent space. To extract new relations in various domains more effectively, we propose to learn more interpretable and robust prototypes by learning from prior knowledge and intrinsic semantics of relations. We improve the prototype representation of relations more efficiently by using prior knowledge to explore the connections between relations. The geometric interpretability of the prototype is improved by making the classification margins between sentence embedding clearer through contrastive learning. Besides, for better-extracting relations in different domains, using a cross-domain approach makes the generation process of the prototype take into account the gap between other domains, which makes the prototype more robust. The experimental results on the benchmark FewRel dataset demonstrate the advantages of the proposed method over some state-of-the-art methods.

1 INTRODUCTION

Relation extraction aims to automatically identify the relations between entities in sentences, which plays a vital role in machine reading comprehension. Relation extraction is often regarded as a multi-classification task and solved by supervised learning methods Kate & Mooney (2010); Riedel et al. (2010). Especially, deep learning methods have achieved impressive performance on this kind of task. The finetuning-based model BERT proposed in Devlin et al. (2018) shows state-of-the-art performance on many classification tasks. However, these methods work based on a large amount of labeled data. When the labeled data is insufficient, their performance degenerates significantly. Relation extraction is a core issue in many scientific fields (e.g., biomedicine and materials). However, in these specific domains, the annotation cost for some classes imposes restrictions on the generalization of current RE models to new relation concepts efficiently. As for this, few-shot learning has raised great interest.

Few-shot learning methods aim to alleviate the heavy reliance on the large annotated corpus since they can identify the sentences of a novel class by exploiting the pre-trained model and a few labeled examples of the novel class. Currently, metric-based few-shot learning methods make classification by calculating the similarity of query sentences and the prototypes generated by a few labeled support samples. However, these methods only work when the novel classes are in the same domain as the classes employed to train the few-shot learner Gao et al. (2019). In other words, the prototypes generated by few-shot learning methods always fail to generalize to classes in different domains. But it is vital due to the difficulty to construct large labeled datasets in the mentioned scientific domains.

To address this issue, the method to bridge the discrepancy for few-shot relation extraction is of great interest. Existing domain adaptation methods can extract a shared feature representation of multiple

different domains Ganin et al. (2016); Shen et al. (2018); Shi et al. (2018). However, these methods only work when the labels on classes in both the source and target domain are the same Gao et al. (2019).

In summary, when a class lacks a large amount of annotated data and its domain also does not have a large labeled corpus, then the problem cannot be perfectly solved by either few-shot learning or domain adaptation separately. For easy description, we call the above problem the cross-domain few-shot relation extraction problem, which is lacking research in natural language processing.

Cross-domain few-shot learning methods have shown the potential to deal with this problem. Although the domain adaptation method in Wang et al. (2018a) can merge data from different domains in the shared latent space learned by the encoder, the generated prototypes have two limitations: 1) it does not explicitly keep the geometrical structure of the classes in the source domain; 2) it does not explicitly minimize the distance between domains.

In this paper, we tackle the cross-domain few-shot relation extraction problem by improving the representation of the prototype. The core observation of this paper is that the distributions of sentence embedding in different classes and domains are significantly different. Consequently, when the metric function used to measure the similarity of the query sentence and prototype may overfit to specific distribution in the training stage and result in a decrease in performance on unseen relations in different domains. To address this issue, we propose to learn more interpretable and robust prototypes by learning from prior knowledge of the connection of relations and intrinsic semantics of relations. We make the prototype more efficient by exploring the connections between relations and increasing the geometric interpretability of the prototype. Moreover, to improve the robustness of the prototypes in different domains, the connection between different domains is also taken into account when generating the prototypes in the training stage.

The contributions of this paper can be summarized as follows:

- We improve the prototype representation of relations more efficiently by using prior knowledge and contrastive learning.
- A cross-domain approach is applied to utilize the gap between various domains, which makes the prototype more robust when the model is used in different domains.
- The proposed method is evaluated on the Pubmed domain and the Semeval domain. It shows that it can significantly outperform some state-of-the-art methods on cross-domain few-shot relation extraction problems.

2 Related work

In the following, the related few-shot learning methods and domain adaptation methods are reviewed in detail.

2.1 Few-shot Learning

Generally, the few-show learning methods can be divided into three categories Munkhdalai & Yu (2017): (1) data-based methods, (2) algorithm-based methods, (3) metric-based methods.

Data-based methods augment the data with prior knowledge to overcome the difficulty of insufficient data Wu et al. (2018); Gao et al. (2018); Cong et al. (2020). For example, Cong et al. Cong et al. (2020) assign pseudo-labels to unlabeled samples for training. It works on cross-domain classification tasks when BERT aligns the features extracted from the source sentence and the target sentence. However, it is time-consuming and requires extra space to train the model.

Algorithm-based methods use prior knowledge to search for an effective initial solution for multiple tasks simultaneously, which makes it easy to adapt to new tasks Finn et al. (2017); Yoo et al. (2018). For example, the model trained by MAML Finn et al. (2017) can work well on new tasks after fine-tuning. Although these methods perform well on many tasks, they cannot work well on the cross-domain relation extraction tasks Gao et al. (2019), as they fail to reduce the discrepancy of different domains.

Metric-based methods learn an encoder based on a metric to refine the sentence embedding in the latent space such that the learned latent space can generalize to novel relations with few labeled samples in the same domain Vinyals et al. (2016); Snell et al. (2017); Triantafillou et al. (2017); Soares et al. (2019). For example, the prototype network Snell et al. (2017) and the matching net Wang et al. (2018b) use Euclidean distance between sentence embedding and relation prototype to identify the relation of the sentence. Generally, these metric-based methods extract the relation of the sentence based on the prototype of the relations, and the prototype is determined by the embedding of labeled sentences in the corresponding relation. The sentences are embedded by a learned encoder. However, the learned encoder in these methods does not explicitly keep the geometric structure of the classes in the latent space. Moreover, they also can not merge different domains with significant discrepancies. Therefore, these methods usually have a good performance on relation extraction tasks with insufficient labeled data only when the tasks belong to the same domain.

2.2 Domain Adaptation

Domain adaption studies how to benefit from different but related domains, and it is employed to deal with various tasks in computer vision Yang et al. (2021); Zhao et al. (2020) and natural language processing Shen et al. (2018); Glorot et al. (2011); Nguyen & Grishman (2014). Unfortunately, some existing domain adaptation methods Nguyen et al. (2015) do not be suitable for our scenario since they require a large number of labeled samples in the target domain in the training process. Although other methods do not require labeled data of the target domain in the training stage, they require different domains to have the same labels, such as comments on laptops and restaurants Fu et al. (2017); Shen et al. (2018); Goodfellow et al. (2014); Shi et al. (2018); Goodfellow et al. (2014); Li et al. (2018); Shi et al. (2018). Therefore, these methods perform well for relation extraction in the target domain only if the target and source domains are highly related. In other words, existing domain adaptation cannot obtain good results for relation extraction if there are non-overlapping relations in the target domain and source domain.

3 Methods

Our key purpose is to improve the generalization ability of few-shot relation extraction models to arbitrary unseen domains by improving the representation of prototypes. There are two domains in the cross-domain few-shot relation extraction problem: the source domain and the target domain. We assume that 1) the source domain and the target domain are significantly different; 2) the labels (relations) on the source domain and target domain are different; 3) there are only a few labeled samples in the target domain. To address the problem, prior knowledge is utilized to explore the connection between different relations in the source domain. And contrastive learning method is also employed to improve the geometric interpretability of the generated prototype. To bridge the gap between these domains, Wasserstein distance is used to modify the representation of prototypes.

The structure of the proposed method is illustrated in Fig. 1, which mainly includes three phases, namely, the learning phase, the adaptation phase, and the prediction phase. In the following, we introduce them one by one.

3.1 Learning phase

The learning phase is to learn an encoder to map the input sentence into the latent space. This paper adopts BERT Devlin et al. (2018) as the encoder. All available data of the source domain and the target domain is used to train the encoder $Enc(\cdot)$. $\mathcal{D}_{\mathrm{S}}$ and $\mathcal{R}_{\mathrm{S}}$ denotes the sentence set and corresponding relation set of the source domain. $\mathcal{R}_{\mathrm{S}}$ includes all different relations in the source domain. $\mathcal{D}_{\mathrm{T}} = \{\mathcal{D}_{\mathrm{LT}}, \mathcal{D}_{\mathrm{UT}}\}$, including the labeled sentence set $\mathcal{D}_{\mathrm{LT}}$ and unlabeled sentence set $\mathcal{D}_{\mathrm{UT}}$, is the sample set of the target domain. The corresponding relation set of $\mathcal{D}_{\mathrm{LT}}$ is denoted by $\mathcal{R}_{\mathrm{LT}}$.

In order to allow the encoder to extract more interpretable prototypes that can be used to improve the relational extraction accuracy and generalizability, this paper proposes to use two loss functions $\mathcal{L}(\theta_{\mathrm{E}})$ and $\mathcal{L}_{\mathrm{adv}}(\theta_{\mathrm{E}})$ for this purpose. The representation loss $\mathcal{L}(\theta_{\mathrm{E}})$ is to make the encoder not only extract the relation of the source domain with prior knowledge but also improve the geometric

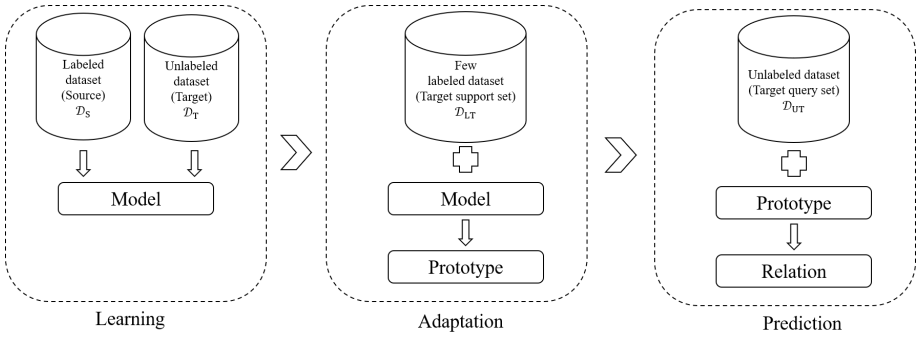

Figure 1: The structure of the proposed method

interpretability of the prototypes in the source domain. And the adversary loss $\mathcal{L}_{\mathrm{adv}}(\theta_{\mathrm{E}})$ is to modify the representation of prototypes while taking the domain discrepancy into consideration.

The representation loss $\mathcal{L}(\theta_{\mathrm{E}})$ is defined as follows.

$$\mathcal{L}(\theta_{\mathrm{E}}) = \mathcal{L}_{\mathrm{cls}} + \rho\mathcal{L}_{\mathrm{con}}, \tag{1}$$

where $\mathcal{L}_{\mathrm{cls}}$ is the cross entropy loss, $\mathcal{L}_{\mathrm{con}}$ is the proposed contrastive loss, and ρ is a hyperparameter, and it is set to 0.2 based on some preliminary experiments.

Like the commonly used few-shot learning methods Gao et al. (2019), the support set $\mathcal{S}$ and query set $\mathcal{Q}$ are randomly selected from the source domain dataset to train the encoder in each training iteration. The support set $\mathcal{S}$ includes N relations, and each relation includes K sentences. The relation set of the support set is denoted as $\mathcal{R}_{\mathcal{S}} = \{r_s | s \in \mathcal{S}\}$. The query set $\mathcal{Q}$ includes the same N relations as the support set, and each relation includes Q sentences.

The prototype $\boldsymbol{v}_{r_i}$ plays a vital role to extract relation r_i. In the initialization, $\boldsymbol{v}_{r_i}, i = 1, \ldots, |\mathcal{R}|$ is defined as follows.

$$\boldsymbol{v}_{r_i} = \boldsymbol{m}_{r_i} + \boldsymbol{h}_{r_i} - \boldsymbol{m}, \tag{2}$$

where $\boldsymbol{m}_{r_i}$ is the mean of the embedding set $\{\boldsymbol{x}_s | s \in \mathcal{D}_S, r_s = r_i\}$; $\boldsymbol{h}_{r_i}$ is the representation of the relation r_i, which is extracted by GNN from the prior knowledge $\mathcal{G} = (\mathcal{R}, \mathcal{W})$. $\mathcal{G} = (\mathcal{R}, \mathcal{W})$ denotes the global relation graph of the source domain, where $\mathcal{R}$ includes all different relations in the source domain, and $\mathcal{W}$ consists of the link weight between relations; $\boldsymbol{m}$ is the mean of the embedding of all sentences (i.e., $\{\boldsymbol{x}_s | s \in \mathcal{D}_S\}$) in the source domain. The details of the calculation for the initial $\boldsymbol{v}_{r_i}$ can refer to Qu et al. (2019).

In the learning phase, the encoder is learned iteratively. In each iteration, a support set $\mathcal{S}$ and a query set $\mathcal{Q}$ are randomly chosen from the source domain dataset to learn the encoder. Similarly, the prototype of the relation is also updated set by set based on the Bayesian model as follows Qu et al. (2020).

$$\boldsymbol{v}_{\mathcal{R}'} \leftarrow \boldsymbol{v}_{\mathcal{R}'} + \frac{\varepsilon}{2}\nabla_{\boldsymbol{v}_{\mathcal{R}'}} \log p(\boldsymbol{v}_{\mathcal{R}'} | \mathcal{X}_{\mathcal{S}}, \mathcal{R}_{\mathcal{S}}, \mathcal{G}) + \sqrt{\varepsilon}\widehat{z}, \tag{3}$$

where $\mathcal{R}'$ denotes the relations sampled for the support set S; $\widehat{z}$ is a random noise from the standard Gaussian distribution; ε is a hyperparameter, and it is set to 0.1 based on some preliminary experiments.

Based on the chain rule, the $p(\mathcal{R}_{\mathcal{S}} | \boldsymbol{x}_{\mathcal{S}}, \boldsymbol{v}_{\mathcal{R}'})$ in Eq. (3) can be calculated as follows.

$$p(\boldsymbol{v}_{\mathcal{R}'} | \mathcal{X}_{\mathcal{S}}, \mathcal{R}_{\mathcal{S}}, \mathcal{G}) \propto p(\mathcal{R}_{\mathcal{S}} | \mathcal{X}_{\mathcal{S}}, \boldsymbol{v}_{\mathcal{R}'}) p(\boldsymbol{v}_{\mathcal{R}'} | \mathcal{G}), \tag{4}$$

where the $p(\boldsymbol{v}_{\mathcal{R}'} | \mathcal{G})$ can be seen as the prior distribution of $\boldsymbol{v}_{\mathcal{R}'}$ and $p(\mathcal{R}_{\mathcal{S}} | \mathcal{X}_{\mathcal{S}}, \boldsymbol{v}_{\mathcal{R}'})$ is the conditional probability of the relation of the sentence in the support set.

The prior distribution $p(\boldsymbol{v}_{\mathcal{R}'} | \mathcal{G})$ of the prototype is parameterized as follows.

$$p(\boldsymbol{v}_{\mathcal{R}'} | \mathcal{G}) = \prod_{r \in \mathcal{R}'} p(\boldsymbol{v}_r | \boldsymbol{h}_r), \tag{5}$$

where $\boldsymbol{h}_r$ is the prototype extracted from the global relation graph $\mathcal{G} = (\mathcal{R}, \mathcal{W})$ Qu et al. (2019).

The conditional probability of the relation of the support set $p(\mathcal{R}_\mathcal{S}|\mathcal{X}_\mathcal{S}, \boldsymbol{v}_{\mathcal{R}'})$ is estimated as follows.

$$\begin{aligned} p(\mathcal{R}_\mathcal{S}|\mathcal{X}_\mathcal{S}, \boldsymbol{v}_{\mathcal{R}'}) &= \prod_{s\in\mathcal{S}} p(r_s|\boldsymbol{x}_s, \boldsymbol{v}_{\mathcal{R}'}) \\ &= \prod_{s\in\mathcal{S}} \prod_{r\in\mathcal{R}'} \frac{\exp(\boldsymbol{x}_s \cdot \boldsymbol{v}_r)}{\sum_{r'\in\mathcal{R}'} \exp(\boldsymbol{x}_s \cdot \boldsymbol{v}_{r'})}. \end{aligned} \tag{6}$$

The prior knowledge is used to modify the representation of prototypes by considering the connection between relations. To explicitly maintain the geometric structure of the relations in the source domain and increase the intrinsic semantics of relations, we introduce a contrastive loss $\mathcal{L}_{\mathrm{con}}$ to deal with this issue for getting more interpretable and robust prototypes for more accurate target domain relation extraction. The contrastive loss in Eq. (1) is defined as follows.

$$\mathcal{L}_{\mathrm{con}} = \mathcal{L}_{\mathrm{S2S}} + \mathcal{L}_{\mathrm{S2V}}, \tag{7}$$

where the $\mathcal{L}_{\mathrm{S2S}}$ means the distance between sentence embedding and the $\mathcal{L}_{\mathrm{S2V}}$ is the distance between sentence embedding and the prototype. By using this loss, we hope the learned encoder can: 1) minimize the distance between sentences in the same class. 2) minimize the distance between the embedding of sentences and their prototypes and maximize the distance between the embedding of sentences and other prototypes.

To minimize the intraclass distance between the embedding of sentences, $\mathcal{L}_{\mathrm{S2S}}$ is defined as follows Soares et al. (2019); Ding et al. (2021).

$$\mathcal{L}_{\mathrm{S2S}} = \frac{1}{\mathrm{N}^2} \sum_{i,j} \frac{\exp(\delta(\boldsymbol{x}_i, \boldsymbol{x}_j))}{\sum_{j'} \exp((1 - \delta(\boldsymbol{x}_i, \boldsymbol{x}_{j'}) d(\boldsymbol{x}_i, \boldsymbol{x}_{j'})))}, \tag{8}$$

where $\boldsymbol{x}_i$ is the embedding of sentence $i \in S$, and

$$\delta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} 1 & r_i = r_j \\ 0 & \text{Otherwise} \end{cases}, \tag{9}$$

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{1 + \exp(\frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|} \cdot \frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|})}, \tag{10}$$

where r_i denotes the relation of sentence $\boldsymbol{x}_i$ in the support set; $d(\cdot, \cdot)$ Soares et al. (2019) denotes the distance between vectors (i.e., the similarity between different vectors).

To minimize the distance between the embedding of sentences and their prototypes and maximize the distance between the embedding of sentences and other prototypes, $\mathcal{L}_{\mathrm{S2V}}$ is defined as follows.

$$\mathcal{L}_{\mathrm{S2V}} = \frac{1}{\mathrm{N}^2} \sum_{r\in\mathcal{R}_\mathcal{S}} \sum_{i=1}^{\mathrm{N*K}} \log(\hat{d}(\boldsymbol{v}_r, \boldsymbol{x}_i)), \tag{11}$$

where

$$\hat{d}(\boldsymbol{v}_r, \boldsymbol{x}_i) = \begin{cases} d(\boldsymbol{v}_r, \boldsymbol{x}_i) & r_i = r \\ 1 - d(\boldsymbol{v}_r, \boldsymbol{x}_i) & \text{Otherwise} \end{cases}. \tag{12}$$

We enable the encoder to extract relations in the source domain more effectively by minimizing $\mathcal{L}_{\theta_{\mathrm{E}}}$. Meanwhile, the accuracy of relation extraction in the target domain is improved. However, it still can not perform well enough when adapting to domains with large discrepancies. To deal with this issue, an adversarial loss $\mathcal{L}_{\mathrm{adv}}$ is proposed to encourage the sentences embedding in different domains as close as possible in the shared latent space so that the prototypes can be modified during training. The adversarial loss $\mathcal{L}_{\mathrm{adv}}(\theta_{\mathrm{E}})$ is defined as follows.

$$\mathcal{L}_{\mathrm{adv}} = \mathrm{Wd}(\mathcal{B}_{\mathrm{Source}}, \tilde{\mathcal{B}}_{\mathrm{Target}}), \tag{13}$$

where $\mathcal{B}_{\mathrm{Source}} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{\mathrm{batch\_size}}\}$ and $\tilde{\mathcal{B}}_{\mathrm{Target}} = \{\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_{\mathrm{batch\_size}}\}$ are minibatch of the sentence embedding in the source domain and target domain, respectively. $\mathrm{Wd}(\cdot, \cdot)$ denotes the wasserstein distance of two subsets.

Algorithm 1 Training for Cross Domain Few-Shot Relation Extraction

Input: Data from source domain and target domain; Global relation graph $\mathcal{G}$ of the source domain; Number of relations in the support set and query set N; Number of sentence(s) in the source domain K; Number of sentence(s) in the query domain Q; Number of epoch E.
Output: The parameter of the encoder θ_{E}
Initialization: $\mathcal{S} = \emptyset, \mathcal{Q} = \emptyset$, the prototypes $v_{\mathcal{R}}$ initialized by Eq. (2).

for $epoch = 1, \cdots, E$ **do**
  Randomly sample N relations $\boldsymbol{v}_{\mathcal{R}'} = \{r_1, \ldots, r_{\mathrm{N}}\}$ in the source domain
  for $j = 1, \ldots, \mathrm{N}$ **do**
    $\mathcal{S} \cup$ SampleSentences($\boldsymbol{x}_i, r_j$), $i = 1, 2, ..., \mathrm{K}$
    $\mathcal{Q} \cup$ SampleSentences($\boldsymbol{x}_i, r_j$), $i = 1, 2, ..., \mathrm{Q}$
  end for
  Update prototype $\boldsymbol{v}_{\mathcal{R}'}$ as Eq. (3).
  Compute representation loss $\mathcal{L}(\theta_{\mathrm{E}})$ by Eq. (1).
  $\theta_{\mathrm{E}} \leftarrow \mathrm{Adam}(\theta_{\mathrm{E}}, \nabla\mathcal{L}(\theta_{\mathrm{E}}))$
  Extract sentence embedding in the support set sampled in the source domain $\mathcal{B}_{\mathrm{Source}}$ and a minibatch of sentence embedding in the target domain $\mathcal{B}_{\mathrm{Target}}$.
  Compute adversarial loss $\mathcal{L}_{\mathrm{adv}}(\theta_{\mathrm{E}})$ by Eq. (13).
  $\theta_{\mathrm{E}} \leftarrow \mathrm{Adam}(\theta_{\mathrm{E}}, \nabla\mathcal{L}_{\mathrm{adv}}(\theta_{\mathrm{E}})$
end for

To reasonably minimize the discrepancy between the source and target domain, the Wasserstein distance (also known as Earth Moving Distance) Eq. 14 is used here Cuturi (2013). The data in both domains follow a discrete probability distribution. These distributions are regarded as quality point scattered across the latent space.

$$\mathrm{Wd}_{M,\alpha}(s,t) := min_{P \in U_\alpha(s,t)} < P, M >, \tag{14}$$

where s and t denote the distribution of the representation of sentences in source $\mathcal{B}_{\mathrm{Source}}$ and target domain $\tilde{\mathcal{B}}_{\mathrm{Target}}$, respectively. P is a joint distribution of source and target domain, which is in the set of $U_\alpha(s,t)$. $M \in \mathbb{R}^{|\mathcal{B}_{\mathrm{Source}}|} \times \mathbb{R}^{|\tilde{\mathcal{B}}_{\mathrm{Target}}|}$ denotes the cost from the source domain to the target domain, where each element in the matrix is computed by a distance metric $M_{ij} = |\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_j|^2$.

Compared with other methods, such as commonly used Kullback-Leibler (KL) divergence, the Wasserstein distance take the structure of the latent space into consideration. Thus, the Wasserstein distance is able to maintain the previous geometric structure while the KL divergence cannot obtain the same performance. The similarity of data with different distributions in the same latent space may not be accurately measured by KL divergence. As the KL divergence between different data distributions may be the same, which cannot take the geometric structure into consideration, but the Wd distance can avoid this problem.

By using the proposed method, the advantage of using contrastive loss can be enhanced. The geometric structure of the source domain will be useful for the classification of the target domain. Therefore, the representation of the sentences will gain better properties.

Finally, based on the loss $\mathcal{L}(\theta_{\mathrm{E}})$, the parameter θ_{E} of the encoder is updated by the Adam optimizer Kingma & Ba (2015). The pseudo-code of training the encoder is shown in Algorithm 1.

3.2 Adaptation phase

In the adaptation phase, a few labeled samples of the target domain is used to generate the prototype of the relations in the target domain based on the learned encoder. We assume that we have a labeled support set $\hat{\mathcal{S}}$ and an unlabeled query set $\hat{\mathcal{Q}}$ in the target domain. The support set $\hat{\mathcal{S}}$ includes $\hat{\mathrm{N}}$ relations, and each relation has $\hat{\mathrm{K}}$ sentences. The query set $\hat{\mathcal{Q}}$ includes some unlabeled sentences. Clearly, the prototypes $\boldsymbol{v}_{\hat{r}}$ generated as follows:

$$\boldsymbol{v}_{\hat{r}} = \frac{1}{\hat{K}} \sum_{i=1}^{K} \hat{\boldsymbol{x}}_i \mathbb{I}(i, r), \tag{15}$$

where $\hat{\boldsymbol{x}}_i$ is the embedding of the sentence i in the support set generated by the learned encoder, and $\mathbb{I}(i, r)$ is an indicator function, defined as

$$\mathbb{I}(i, r) = \begin{cases} 1 & r_i = \hat{r} \\ 0 & \text{Otherwise} \end{cases} . \tag{16}$$

3.3 PREDICTION PHASE

The prediction phase is to predict the relation of the sentence of the query set $\hat{\mathcal{Q}}$ in the target domain. Based on the prototype of the relation in the query set, the relation of a sentence q is determined as

$$r_q = \text{argmax}_r \frac{\hat{\boldsymbol{x}}_q \cdot \hat{\boldsymbol{v}}_r}{\sum_{i=1}^{\text{N}} \hat{\boldsymbol{x}}_q \cdot \boldsymbol{v}_{r_i}}. \tag{17}$$

4 EXPERIMENTS

In this section, we conduct experiments on one benchmark dataset to evaluate our proposed approach. We make a comprehensive analysis of our approach and compare it with state-of-the-art approaches.

4.1 DATA

In the experimental study, the FewRel dataset Gao et al. (2019) is chosen, which is a widely used benchmark for few-shot relation extraction. It contains data from four different domains, including Wikipedia, SemEval-2010 task 8, NYT, and Pubmed. For our experiment setting, we use 44,800 sentences (64 classes and 700 sentences per class) from Wikipedia as the training set and 11,200 sentences (16 classes and 700 sentences per class) from Wikipedia as the validation set. And we use 1,000 sentences (10 classes and 100 sentences per class) from Pubmed. Also, we use Semeval as the testing set to conduct another experiment. The Wikipedia data serves as the source domain, while the Pubmed and Semeval data are the target domains. There are no overlapping sentences between training, validation, and testing sets.

Beyond that, a global knowledge graph that consists of 828 unique relations in the source domain serves as the prior knowledge. The embedding of each relation in the graph has been processed by TransE algorithm Bordes et al. (2013). Then the graph is constructed as a 10-nearest neighbor graph as the final global relation graph in the source domain $\mathcal{G}$. The graph only contains relations in Wikipedia dataset (source domain), which can not be used to train the model on other datasets.

4.2 COMPARISON AND ANALYSIS

We choose the following methods for comparison.

Proto Snell et al. (2017): The algorithm of the prototype network. A few-shot relation extraction method that extracts relations by measuring the distance between the sentence embedding and the prototype.

Proto+adv Gao et al. (2019): The Proto algorithm uses a discriminator to adjust the source and target domains.

MTB Soares et al. (2019): The algorithm, called Matching The Blanks, builds task agnostic relation representations solely from the entity-linked text.

GNN Garcia & Bruna (2017): The algorithm uses Graph Neural Network (GNN) to predict the relation.

MAML Finn et al. (2017): The algorithm, called Model-Agnostic Meta-Learning, solves few-shot learning problems by meta-learning method.

Siamese Koch et al. (2015): The algorithm uses temporal CNN and an attention mechanism for few-shot learning.

DaFeC Cong et al. (2020): The algorithm improves domain adaptation performance for few-shot classification via clustering.

REGRAD Qu et al. (2020): The algorithm completes the few-shot relation extraction task via Bayesian meta-learning on the relation graph.

REGRAD+adv Qu et al. (2020): The algorithm adds an adversarial part to the REGRAD model.

Alg	**5-way 1-shot**	**5-way 5-shot**	**10-way 1-shot**	**10-way 5-shot**
Proto	66.22	77.47	49.77	65.63
Proto+adv	41.09	67.26	28.32	40.01
MTB	56.42	68.28	40.74	54.56
GNN	36.44	37.19	26.00	28.07
Siamese	66.58	78.69	52.30	64.45
MAML	66.62	78.53	51.90	65.57
DaFeC	30.21	30.51	15.17	17.27
Regrad	71.70	80.74	61.66	74.06
Regrad+adv	65.10	71.61	56.44	56.71
ours	**76.30**	**84.71**	**67.87**	**75.84**

Table 1: **Results of the cross-domain few-shot relation extraction on the Pubmed dataset.** We run all the algorithms on the same conditions.

Alg	**5-way 1-shot**	**5-way 5-shot**	**10-way 1-shot**	**10-way 5-shot**
Proto	41.39	59.51	27.62	42.96
Proto+adv	26.96	48.06	13.15	28.19
MTB	34.03	47.90	21.40	30.57
GNN	32.13	37.12	14.71	17.92
Siamese	41.67	53.57	28.06	39.52
MAML	42.75	52.87	27.89	43.06
DaFeC	24.72	25.98	11.17	13.37
Regrad	49.56	64.57	36.17	54.10
Regrad+adv	50.71	65.46	38.61	54.56
ours	**53.21**	**67.16**	**39.99**	**57.00**

Table 2: **Results of the cros-domain few-shot relation extraction on the Semeval dataset.** We run all the algorithms on the same conditions.

As there are few studies on the cross-domain few-shot relation extraction task, the state-of-the-art algorithms in the few-shot relation extraction task and few-shot relation extraction algorithm together with the adversarial part Gao et al. (2019); Qu et al. (2020); Cong et al. (2020) are chosen as the baseline in this paper. The Regrad and Regrad+adv algorithms are re-implemented as the paper Qu et al. (2020). The DaFeC algorithm is re-implemented as the paper Cong et al. (2020). Other algorithms are re-implemented by Gao et al. Gao et al. (2019). $\text{Bert}_{\text{base}}$ is used as the encoder to project the sentences into the latent space for all algorithms. Besides, the hyper-parameters used in our method remain the same with the setting of Gao et al. (2019); Qu et al. (2020).

The prediction accuracy on the target domain is used as the criterion to judge the performance of the algorithms. The comparison results are shown in Table 1 and Table 2.

The performance of these baselines is not competitive on the cross-domain few-shot relation extraction task. The results of GNN and DaFeC are less competitive, showing that these methods are less effective to solve the cross-domain few-shot relation extraction task. The methods specifically designed for few-shot tasks, such as prototype network (Proto) and MTB, cannot outperform well on this specific task. Compared with other meta-learning methods, such as MAML and Siamese, our approach can better generalize to different domains. The most competitive algorithm is the Regrad, but the performance on the cross-domain task can not surpass our method. Previous adversarial methods only merge the source and target domains by puzzling the employed discriminator. However, the effectiveness of the method is highly correlated with the dataset and algorithm. In other words, the performance of the algorithm will reduce when applied to some datasets and algorithms.

Our model surpasses state-of-the-art models because our model can ensure a better geometric structure of the latent space. The distance between sentence embedding in the same class is closer, and

Alg	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
Original	71.70	80.74	61.66	**74.06**
With Wd	72.79	80.69	64.55	73.49
With con	73.80	81.02	62.95	73.25
With Wd and con	**76.30**	**84.71**	**67.87**	**75.84**

Table 3: Ablation results of the cross-domain few-shot relation extraction on the Pubmed dataset.

Alg	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
Original	49.56	64.57	36.17	54.10
With Wd	52.19	66.08	39.26	54.79
With con	51.67	65.03	39.42	54.27
With Wd and con	**53.21**	**67.16**	**39.99**	**57.00**

Table 4: Ablation results of the cross-domain few-shot relation extracttion on the Semeval dataset.

the distance in different classes is farther. In addition, when the prototype is the relation representation of a given sentence embedding, the distance between the sentence embedding and the prototype is closer, otherwise, it is farther. Besides, by optimizing the adversarial loss, the distribution of the target domain is as close as possible to the source domain. Thus, the performance is further improved in the cross-domain relation extraction task.

4.3 Ablation Study

In this subsection, we study the impact of contrastive loss and adversarial loss on generalization accuracy. The model only optimizes cross-entropy loss $\mathcal{L}_{\text{cls}}$ is named as original model here. We conduct some ablation study on FewRel dataset, where we compare three variant methods, i.e., original model with $\mathcal{L}_{\text{Wd}}$, with $\mathcal{L}_{\text{con}}$ and with both of the loss. The results are presented in Table 3 and Table 4.

First, we find that the contrastive loss effectively improves the performance of the target domain by utilizing the geometric structure of the latent space. Moreover, the adversarial loss further improves the performance of the target domain by reducing the discrepancy between the source and target domains. The observation shows that combining both of the loss can help the method solve the cross-domain few-shot relation extraction problem well.

5 Conclusion

In this paper, we have proposed a novel method by integrating the method of few-shot learning and domain adaptation to solve the cross-domain few-shot relation extraction task. To improve the interpretably of the representation of prototypes, we have designed a representation loss, including a cross-entropy loss and a contrastive loss. Besides, an adversarial loss has been further employed to consider the discrepancy between different domains. Extensive experiments have demonstrated that our method performs better than some existing state-of-the-art relation extraction methods. Moreover, the effectiveness of each used loss also has been validated by experiment.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Neural Information Processing Systems*, pp. 1–9, 2013.

Xin Cong, Bowen Yu, Tingwen Liu, Shiyao Cui, Hengzhu Tang, and Bin Wang. Inductive unsupervised domain adaptation for few-shot classification via clustering. *arXiv preprint arXiv:2006.12816*, 2020.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Neural Information Processing Systems*, pp. 2292–2300, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ning Ding, Xiaobin Wang, Yao Fu, Guangwei Xu, Rui Wang, Pengjun Xie, Ying Shen, Fei Huang, Hai-Tao Zheng, and Rui Zhang. Prototypical representation learning for relation extraction. *arXiv preprint arXiv:2103.11647*, 2021.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017.

Lisheng Fu, Thien Huu Nguyen, Bonan Min, and Ralph Grishman. Domain adaptation for relation extraction with domain adversarial neural network. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 425–429, 2017.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Hang Gao, Zheng Shou, Alireza Zareian, Hanwang Zhang, and Shih-Fu Chang. Low-shot learning via covariance-preserving adversarial augmentation networks. *arXiv preprint arXiv:1810.11730*, 2018.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fewrel 2.0: Towards more challenging few-shot relation classification. *arXiv preprint arXiv:1910.07124*, 2019.

Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*, 2017.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *International Conference on Machine Learning*, 2011.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

Rohit Kate and Raymond Mooney. Joint entity and relation extraction using card-pyramid parsing. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pp. 203–212, 2010.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *International Conference on Machine Learning*, volume 2. Lille, 2015.

Yitong Li, Michael Murias, Samantha Major, Geraldine Dawson, and David E Carlson. Extracting relationships by multi-domain matching. In *Neural Information Processing Systems*, pp. 6799–6810, 2018.

Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *International Conference on Machine Learning*, 2017.

Thien Huu Nguyen and Ralph Grishman. Employing word representations and regularization for domain adaptation of relation extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 68–74, 2014.

Thien Huu Nguyen, Barbara Plank, and Ralph Grishman. Semantic representations for domain adaptation: A case study on the tree kernel-based method for relation extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 635–644, 2015.

Meng Qu, Yoshua Bengio, and Jian Tang. Gmnn: Graph markov neural networks. In *International Conference on Machine Learning*, 2019.

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

Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 148–163. Springer, 2010.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Ge Shi, Chong Feng, Lifu Huang, Boliang Zhang, Heng Ji, Lejian Liao, and He-Yan Huang. Genre separation network with adversarial training for cross-genre relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1018–1023, 2018.

Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*, 2019.

Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. Few-shot learning through an information retrieval lens. *arXiv preprint arXiv:1707.02610*, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016.

Xiaozhi Wang, Xu Han, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Adversarial multi-lingual neural relation extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1156–1166, 2018a.

Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7278–7286, 2018b.

Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5177–5186, 2018.

Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. *arXiv preprint arXiv:2101.06395*, 2021.

Donghyun Yoo, Haoqi Fan, Vishnu Boddeti, and Kris Kitani. Efficient k-shot learning with regularized deep networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Sicheng Zhao, Guangzhi Wang, Shanghang Zhang, Yang Gu, Yaxian Li, Zhichao Song, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source distilling domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12975–12983, 2020.