Learn from Relation Information: Towards Prototype Representation Rectification for Few-Shot Relation Extraction

Yang Liu[♠], Jinpeng Hu[♠], Xiang Wan[♠♢†], Tsung-Hui Chang[♠†]

[♠]Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, Guangdong, China

[♢]Pazhou Lab, Guangzhou, 510330, China

[♠]{yangliu5, jinpenghu}@link.cuhk.edu.cn

[♢]wanxiang@sribd.cn [♠]changtsunghui@cuhk.edu.cn

Abstract

Few-shot Relation Extraction refers to fast adaptation to novel relation classes with few samples through training on the known relation classes. Most existing methods focus on implicitly introducing relation information (i.e., relation label or relation description) to constrain the prototype representation learning, such as contrastive learning, graphs, and specifically designed attentions, which may bring useless and even harmful parameters. Besides, these approaches are limited in handing outlier samples far away from the class center due to the weakly implicit constraint. In this paper, we propose an effective and parameter-less **P**rototype **R**ectification **M**ethod (PRM) to promote few-shot relation extraction, where we utilize a prototype rectification module to rectify original prototypes explicitly by the relation information. Specifically, PRM is composed of two gate mechanisms. One gate decides how much of the original prototype remains, and another one updates the remained prototype with relation information. In doing so, better and stabler global relation information can be captured for guiding prototype representations, and thus PRM can robustly deal with outliers. Moreover, we also extend PRM to both none-of-the-above (NOTA) and domain adaptation scenarios. Experimental results on FewRel 1.0 and 2.0 datasets demonstrate the effectiveness of our proposed method, which achieves state-of-the-art performance.[1][2]

1 Introduction

Relation Extraction (RE) is one of the fundamental natural language processing (NLP) tasks, which aims to detect the relation between two entities contained in a sentence. Most RE models (Distiawan et al., 2019; Li et al., 2019; Jin et al., 2020) require large labeled datasets while constructing such datasets is usually high-costing and time-consuming. Thus, the Few-shot Relation Extraction (FSRE) has become a hot topic to alleviate data scarcity. There are two main steps in FSRE. The model is first trained on collections of few-shot tasks (i.e., meta tasks) sampled from the large-scale data containing disjoint relations and then fast adapted to the unseen relation classes with few samples. Recently, many approaches have been proposed for addressing FSRE problems (Han et al., 2018; Gao et al., 2019b; Qu et al., 2020; Baldini Soares et al., 2019). One of the popular algorithms is the Prototype Network (Snell et al., 2017), which is based on the meta-learning framework (Vilalta and Drissi, 2002; Vanschoren, 2018), and the basic framework used in the paper. Prototype Network generates a prototype representation for each relation class in the meta task with the given instances (generally average instances in each relation class). Then, the distance of query instances and each class prototype are calculated for model train and prediction.

To achieve better performance, many works have integrated the relation information into the model to assist prototype representation learning. TD-proto (Yang et al., 2020) enhanced prototypical network with both relation and entity descriptions. CTEG (Wang et al., 2020) proposed a model that learns to decouple high co-occurrence relations, where two types of external information are added. MapRE (Dong et al., 2021) proposed a framework considering both label-agnostic and label-aware semantic mapping information in pre-training and fine-tuning. HCRP (Han et al., 2021) introduced three modules containing hybrid prototype learning, relation-prototype contrastive learning, and task adaptive focal loss for the model improvement.

[†]Corresponding author.

[1]The code is released at https://github.com/lylylylylyly/PRM-FSRE

[2]Main results in this paper can be found in the CodaLab competition (username is *atry*), which you can get the three competition websites, i.e., FewRel 1.0, FewRel 2.0 (Domain Adaptation), and FewRel 2.0 (NOTA) from https://thunlp.github.io/fewrel.html

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1822 - 1831

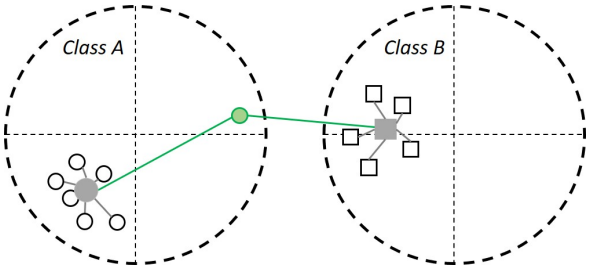

Figure 1: An Illustration of a possible misclassification case. *Class A* and *Class B* represent different relation classes, while circles and squares represent samples of corresponding classes. The green circle represents the sample that needs to be classified.

However, most existing methods mainly utilized instances given in each relation class to obtain the prototype representation (generally to average these instances). Although they implicitly incorporate relation information to constrain the prototype representations learning by contrastive learning, graphs, or attentions, such insufficiently and weakly implicit constraints are limited in dealing with the outlier samples. We provide an example in Fig. 1, where $ClassA$ and $ClassB$ represent two different relation classes; the circles and squares represent the few instances provided by each relation class, and the green circle represents the instance that needs to be predicted. It can be seen that if the provided instances are remote and not "good", the model will tend to classify the instance (green circle) into the $ClassB$. Besides, existing methods also introduced more parameters into the model owing to their specific designs, which is detrimental to FSRE. More parameters mean a more complex model, which increases the overfitting risk on the training set, thereby reducing the generalization ability of the model (Dar et al., 2021).

To address aforementioned issues, this paper proposes a **P**rototype representation **R**ectification **M**ethod, named **PRM**, which focuses on obtaining a better prototype for each relation. Specifically, we propose a prototype rectification module, capable of explicitly utilizing relation information and instances to generate the rectified prototype representations together instead of implicitly using the relation information to guide the generation of the prototype representation. Stated in another way, our model tries to use relation information to rectify the distribution of the original few instances to make it more global and more representative for the overall distribution of the class. PRM transforms the problem of perceiving the class distribution from local instances to perceiving the class distribution from local instances and global information.

In addition, we extend PRM to an advanced version of the existing N-way K-shot setting in few-shot learning (i.e., None-Of-The-Above (NOTA) scenario), where queries could also be none-of-the-above instead of assuming that all query instances belong to the sampled N classes of supports. Although this task brings one more option in classification and is more challenging for the general FSRE model, our model can easily extend to NOTA by introducing an external description "*The relation of the query is not the same as this prototype.*". Experiments on FewRel 1.0 (general scenario) and FewRel 2.0 (NOTA scenario) demonstrate the effectiveness of our proposed method with state-of-the-art results.

The **contribution** of our work mainly lies in three folds:

(1) We introduce the idea of using relation information to rectify prototypes explicitly and propose an effective and parameter-less method, PRM, compared to previous works with always complex modules or networks.

(2) In PRM, a prototype rectification module is utilized, which explicitly utilizes relation information and instances to generate rectified prototypes.

(3) We further extend PRM to the NOTA scenario that is an advanced version of the existing N-way K-shot setting in few-shot learning and then justify the easy transferability of PRM to both NOTA and domain adaptation scenarios.

2 Related Work

Relation Extraction (RE) (Kumar, 2017; Han et al., 2020) is a fundamental task for information extraction, aiming to recognize the relation types that exist between entity pairs in one sentence. The labeling of relations is usually time-consuming and laborious. In addition, in some specific fields, such as the medical field, the available data are few and additional expertise is required. Therefore, the Few-shot Relation Extraction (FSRE) task has attracted more and more attention recently. FSRE aims to fast adapt to unseen relation classes with few samples through training on known relation classes. Garcia and Bruna (2018); Gao et al. (2019b) proposed a large-scale supervised few-shot relation classification dataset, namely, FewRel, and provided the current state-of-the-art results on FewRel, i.e., Proto-BERT (Garcia and Bruna, 2018) and BERT-PAIR. Most subsequent work is evaluated

on FewRel. REGRAB (Qu et al., 2020) proposed to incorporate an external global relation graph based on a Bayesian meta-learning method. Except for relation descriptions, TD-proto (Yang et al., 2020) and ConceptFERE (Yang et al., 2021) also introduced entity descriptions to provide clues for relation prediction and enhancing the prototype network. CTEG (Wang et al., 2020) proposed a model that learns to decouple high co-occurrence relations, where two external information is added. MapRE (Dong et al., 2021) proposed a framework considering both label-agnostic and label-aware semantic mapping information for low resource relation extraction in both pre-training and fine-tuning. HCRP (Han et al., 2021) introduced three modules containing hybrid prototype learning, relation-prototype contrastive learning, and task adaptive focal loss for the model improvement. However, these methods always introduced relation information implicitly, which may introduce more parameters and are limited in dealing with outlier samples. Thus, explicitly rectifying the prototypes with relation representations can be a more effective way to incorporate relation information.

3 Task Definition

We follow a typical few-shot task setting, namely the N-way-K-shot setup, which contains a support set $\mathcal{S}$ and a query set $\mathcal{Q}$. The support set $\mathcal{S}$ includes N novel classes, each with K labeled instances. The query set $\mathcal{Q}$ contains the same N classes as $\mathcal{S}$. And the task is evaluated on the query set $\mathcal{Q}$, trying to predict the relations of instances in $\mathcal{Q}$. What's more, an auxiliary dataset $\mathcal{D}_{base}$ is given, which contains abundant base classes, each with a large number of labeled examples. Note the base classes and novel classes are disjoint with each other. The few-shot learner aims to acquire knowledge from base classes and use the knowledge to recognize novel classes. One popular approach is the meta-learning paradigm, which mimics the few-shot learning settings at the training stage. Specifically, in each training iteration, we randomly select N classes from base classes, each with K instances to form a support set $\mathcal{S} = \{s_k^i; i = 1, ..., N, k = 1, ..., K\}$. Meanwhile, G instances are sampled from the remaining data of the N classes to construct a query set $\mathcal{Q} = \{q_j; j = 1, ..., G\}$. The model is optimized by collections of few-shot tasks sampled from base classes so that it can rapidly adapt to new tasks.

For an FSRE task, each instance consists of a set of samples (x, e, y), where x denotes a natural language sentence, $e = (e_h, e_t)$ indicates a pair of the head entity and tail entity, generally called statements, and y is the relation label. The name and description for each relation are also provided as auxiliary support evidence for relation extraction.

4 Proposed Method

In this section, we present the details of our proposed approach. Figure 2 shows the overall structure, where three colors are used to represent different relation types. The inputs are N-way K-shot tasks (sampled from the auxiliary dataset D_{base}), where each task contains a support set S and a query set Q. Meanwhile, we take the names and descriptions of these N classes (i.e., relations) as inputs as well. All input items share the same sentence encoder. The prototype rectification module utilizes relation representations and the mean value of representations of the given instances, called *Original Prototypes*, to generate the rectified prototypes together. Then, the model calculates the distance between the rectified prototypes and each query for both training and predicting.

4.1 Sentence Encoder

We employ BERT (Devlin et al., 2019) as the encoder to obtain contextualized embeddings of support instances and query instances. Then, the statements of these instances are obtained by concatenating the hidden states corresponding to start tokens of two entity mentions following (Baldini Soares et al., 2019). Denote statements of support instances as $\{\mathbf{S}_k^i \in \mathbb{R}^{2d}; i = 1, ..., N, k = 1, ..., K\}$ (i.e., solid circles in Fig. 2), and statements of query instances as $\{\mathbf{Q}_j \in \mathbb{R}^{2d}; j = 1, ..., G\}$ (i.e., diamonds in Fig. 2), where d denotes the hidden size of BERT output.

For each relation, we concatenate the name and description with a template "name: description", and then feed the sequence into the BERT encoder to obtain relation embeddings. For example, we combine the relation name *"debut participant"* with its description *"participant for whom this is their debut appearance in a series of events"* as the sequence *"debut participant: participant for whom this is their debut appearance in a series of events"*. In more detail, we concatenate the hidden states corresponding to the [*CLS*] token and the mean value of hidden states of all tokens to obtain the

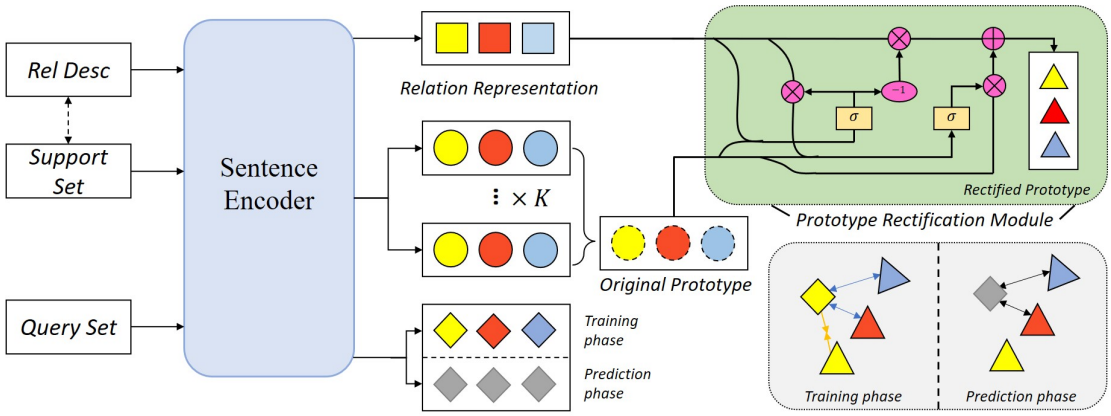

Figure 2: The overall structure of our proposed method. The same sentence encoder is used to encode relations, support set and query set. The relations and *Original Prototypes* are fed into the prototype recification module together to obtain *Rectified Prototypes*.

relation representations $\{R^i \in \mathbb{R}^{2d}, i = 1, ..., N\}$ (i.e., squares in Fig. 2).

4.2 Prototype Rectification Module

Original Prototypes The FSRE task based on the Prototype Network (Snell et al., 2017) paradigm generally first obtains the statement representations of the K instances in each relation class (a total of N relations) in the support set with BERT encoding, and then takes the average of the K statements representations to obtain relation prototypes. We call the prototype obtained in this way *Original Prototype* (i.e., dotted circles in Fig. 2) and denote as $\{P^i_{ori} \in \mathbb{R}^{2d}, i = 1, ..., N\}$. Since *Original Prototypes* is completely obtained from the K instances given for each relation type in the support set, once the K instances are not "good" enough and too far from the true class center, it will cause the model to make wrong predictions.

Rectified Prototypes The name and description of the relationship class (we will refer to them collectively as "relations" in the paper for simplification) are the naive pieces of information that can characterize the overall class distribution and are easily accessible for the FSRE task. Based on the above facts, we propose to utilize relations to rectify *Original Prototypes*, so that the *Rectified Prototypes* (i.e., solid triangles in Fig. 2) contain both the global distribution information in relations and the local distribution information of the K specific instances given for each relation class.

Inspired by GRU (Cho et al., 2014), multiple gate mechanisms are used to control how much *Original Prototypes* are retained and how much relations information is introduced for generating *Rectified Prototypes* together. Firstly, we obtain how much relation information is introduced and how much relation information should be replaced by *Original Prototypes* through a gate mechanism performing on relations and *Original Prototypes*:

$$r^i = \sigma(W_r \cdot [R^i, P^i_{ori}] + b_r) \tag{1}$$

$$\begin{aligned} R^i_{remain} &= ([1] - r^i) \times R^i \\ R^i_{replace} &= r^i \times R^i \end{aligned} \tag{2}$$

where $i = 1, ..., N$; $[\cdot, \cdot]$ denotes concatenation operation; r^i is a weight value for the relation class i. Then another gate is used to control how much information of *Original Prototypes* needed for the *Rectified Prototypes* generation.

$$\begin{aligned} p^i &= \sigma(W_p \cdot [R^i_{replace}, P^i_{ori}] + b_p) \\ P^i_{ori-remain} &= p^i \times P^i, i = 1, \ldots, N \end{aligned} \tag{3}$$

Finally, we obtain *Rectified Prototypes* by the summation of R^i_{remain} and $P^i_{ori-remain}$:

$$P^i_{rec} = R^i_{remain} + P^i_{ori-remain} \tag{4}$$

where $i = 1, \ldots, N$; $P^i_{rec} \in \mathbb{R}^{2d}$. Note that All the representations used above belong to the $\mathbb{R}^{2d}$ feature space.

4.3 Training Objective

The model uses the **vector dot product** way to calculate the distance between the query instance Q and each *Rectified Prototypes* $\{P^i_{rec}, i = 1, ..., N\}$, and then feed the distance into cross entropy loss to form the training loss, which is similar to the contrastive loss. Finally, the training loss $\mathcal{L}$ is defined

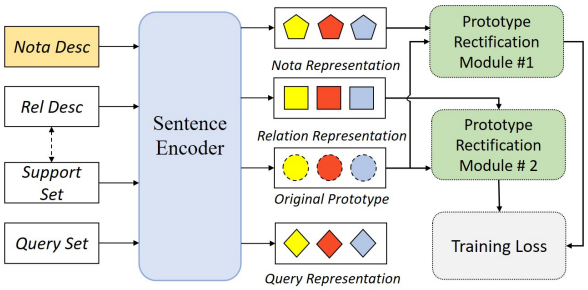

Figure 3: PRM on NOTA scenario. An additional NOTA description is utilized.

via *Rectified Prototypes* and query instances:

$$\mathcal{L} = -\sum_{j=1}^{G} log \frac{exp(P_{rec}^{i} \cdot Q_j)}{\sum_{i=1}^{N} exp(P_{rec}^{i} \cdot Q_j)} \quad (5)$$

In the prediction stage, the model calculates the distance between *Rectified Prototypes* and query instances again and selects the relation class with the shortest distance as the prediction result.

4.4 NOTA Scenario

To verify the effectiveness of the proposed PRM, we further extend it to the none-of-the-above (NOTA) scenario. NOTA is an advanced version of the existing N-way K-shot setting. The original N-way K-shot setting samples N classes, as well as K supporting instances and several queries from each class for each test batch, assuming that all queries belong to the sampled N classes. However, in few-shot NOTA, queries could also be none-of-the-above (NOTA), which brings one more option in classification (i.e., $(N+1)$-way K-shot) and challenges existing few-shot methods. The difficulty in solving the NOTA scenario based on the proposed PRM is how to obtain the representation or distance with queries of the one more class, since this additional class information does not refer to a specific relation class, it represents a meaning that the query instance does not belong to any relation class in support instances.

We introduce an external description to describe the NOTA class, that is *"The relation of the query is not the same as this prototype."*. The NOTA description also shares the same sentence encoder as relations and instances. However, the NOTA description does not share the same prototype rectification module with relations. Specifically, we feed the *Original Prototypes* and NOTA description representations into another prototype rectification module and get prototypes containing NOTA information for each relation in support set, named as *NOTA Prototypes*, $\{P_{nota}^{i}, i = 1, ..., N\}$. Then, we calculate the vector dot product of each query instance to *NOTA Prototypes* as the distance and take the smallest distance value.

$$D_{nota}^{j} = min\{P_{nota}^{i} \cdot Q_j, i = 1, ..., N\} \quad (6)$$

where j denotes the index of query instances and i denotes the index of relations.

Finally, D_{nota}^{j} and distances between *Rectified Prototype* and query instances are fed into cross entropy loss together.

$$\mathcal{L} = -\sum_{j=1}^{G} log \frac{exp(P_{rec}^{i} \cdot Q_j) \ \ or \ \ D_{nota}^{j}}{\sum_{i=1}^{N} exp(P_{rec}^{i} \cdot Q_j)} \quad (7)$$

If the true label is NOTA, then the numerator in the formula above is D_{nota}^{j} during training.

5 Experiments

5.1 Dataset

Our proposed approach is evaluated on the commonly used large-scale FSRE dataset FewRel 1.0 and FewRel 2.0 (Han et al., 2018; Gao et al., 2019b), which are constructed from Wikipedia and consist of 100 relations, each with 700 labeled instances. The average number of tokens in each sentence instance is 24.99, and there are 124,577 unique tokens in total. In addition, the name and description of each relation are also given, providing additional interpretability for each relation. FewRel 2.0 with none-of-the-above setting is a more challenging task to detect none-of-the-above (NOTA) relations for queries. Moreover, FewRel 2.0 with domain adaptation setting is used in the transferability analysis in Section 7.2 that is trained on Wikipedia domain but tested on a different biomedical domain. Only the names of relation labels are given but their descriptions are not available, which makes the task more challenging. Our experiments follow the splits used in official benchmarks with 64 base classes for training, 16 classes for validation, and 20 novel classes for testing.

5.2 Implementation Details

Evaluation N-way K-shot (N-w-K-s or NwKs) is commonly used to simulate the distribution of FewRel in different situations, where N and K denote the number of classes and samples from each class, respectively. In the N-w-K-s scenario, accuracy is used as the performance metric. To be noted, consistent with the official evaluation scripts,

Encoder	Model	5-w-1-s	5-w-5-s	10-w-1-s	10-w-5-s
CNN	Proto-HATT	72.65 / 74.52	86.15 / 88.40	60.13 / 62.38	76.20 / 80.45
	MLMAN	75.01 / — —	87.09 / 90.12	62.48 / — —	77.50 / 83.05
BERT	BERT-PAIR	85.66 / 88.32	89.48 / 93.22	76.84 / 80.63	81.76 / 87.02
	Proto-BERT*	84.77 / 89.33	89.54 / 94.13	76.85 / 83.41	83.42 / 90.25
	REGRAB	87.95 / 90.30	92.54 / 94.25	80.26 / 84.09	86.72 / 89.93
	CTEG	84.72 / 88.11	92.52 / 95.25	76.01 / 81.29	84.89 / 91.33
	ConceptFERE	— — / 89.21	— — / 90.34	— — / 75.72	— — / 81.82
	HCRP (BERT)	90.90 / 93.76	93.22 / 95.66	84.11 / 89.95	87.79 / 92.10
	PRM (BERT)	91.08 / **94.22**	93.72 / **96.51**	84.67 / **91.42**	88.82 / **92.79**
	MTB	— — / 91.10	— — / 95.40	— — / 84.30	— — / 91.80
	CP	— — / 95.10	— — / 97.10	— — / 91.20	— — / 94.70
	MapRE	— — / 95.73	— — / 97.84	— — / 93.18	— — / 95.64
	HCRP (CP)	94.10 / 96.42	96.05 / 97.96	89.13 / 93.97	93.10 / 96.46
	PRM (CP)	95.10 / **96.64**	97.11 / **98.05**	91.12 / **94.55**	94.90 / **96.55**
	Δ (BERT)	+4.89	+2.38	+8.01	+2.54
	Δ (CP)	+1.54	+0.95	+3.35	+1.85

Table 1: Experimental results of FSRE on FewRel 1.0 validation/test set, where N-w-K-s stands for the abbreviation of N-way-K-shot. The table divides the method with BERT as the encoder into two parts, from top to bottom including approaches with the original BERT, and approaches with additional pre-training on BERT. Note that $*$ represents the results of our implementation, others are obtained from results reported by papers or CodaLab.

Model	5w1s (0.15)	5w5s (0.15)	5w1s (0.5)	5w5s (0.5)	Aver.
BERT-PAIR	77.67	84.19	80.31	86.06	82.06
MNAV♮	79.06	85.52	81.69	87.74	83.50
lfc♮	82.61	87.46	80.17	80.84	82.77
PRM (BERT)	83.01	89.30	83.32	85.94	85.39
PRM (CP)	**91.58**	**93.63**	**89.81**	**91.05**	**91.52**

Table 2: Experimental results of FSRE on FewRel 2.0 (NOTA) test set, where 0.15, 0.5 specifies the rate between Q for NOTA and Q for positive, where ♮ denotes the result obtained from *Codalab*.

we select the best model for the test by evaluating our model on randomly sampling 10,000 tasks from validation data. Since the label of the test set of the FewRel is not publicly available, we submit the prediction file of our best model to the official leaderboard in *CodaLab* to obtain the final result on the test set.

Training We use BERT-base-uncased and CP (Wang et al., 2020) as the sentence encoder, where CP is a further pre-trained model based on BERT with contrastive learning. We set the total train iteration number as 30,000, validation iteration number as 1,000, batch size as 4, learning rate as 1e-5 and 5e-6 for BERT and CP respectively.

5.3 Comparable Models

5.3.1 General Scenario

We compare our proposed method with eleven baselines in total. Based on the type of encoder, the comparable models are divided into two types, namely, two CNN-based models and nine BERT-based models. Specifically, CNN-based models include: 1) **Proto-HATT** (Gao et al., 2019a), prototypical networks modified with hybrid attention to focus on the crucial instances and features. 2) **MLMAN** (Ye and Ling, 2019), a multi-level matching and aggregation prototypical network. BERT-based models include: 3) **Proto-BERT** (Garcia and Bruna, 2018), a method that measures the similarity of prototypes and query instances for each relation. 4) **BERT-PAIR** (Gao et al., 2019b), a method that measures the similarity of sentence pairs. 5) **REGRAB** (Qu et al., 2020), a Bayesian meta learning method with an external global relation graph. 6) **CTEG** (Wang et al., 2020), a model that learns to decouple high co-occurrence relations, where two external information is added. 7) **ConceptFERE** (Yang et al., 2021), introducing the inherent concepts of entities to provide clues for relation prediction. 8) **MTB** (Baldini Soares et al., 2019), pre-train with their proposed matching the blank task on top of an existing BERT model. 9) **CP** (Peng et al., 2020), an entity masked con-

Model	5-w -1-s	5-w -5-s	10-w -1-s	10-w -5-s
Proto-BERT	84.77	89.54	76.85	83.42
w/ relation info.				
–Add	89.15	93.11	83.63	87.93
–Concat	80.34	85.11	73.78	80.85
w/ PRM	**91.08**	**93.72**	**84.67**	**88.82**

Table 3: Ablation Study in the validation set. w/ is the abbreviations of with. PRM is the abbreviation of Prototype Rectification Method/Module.

trastive pre-training framework for RE while utilizing prototype networks for fine-tuning on FSRE. 10) **MapRE** (Dong et al., 2021), a framework considering both label-agnostic and label-aware semantic mapping information in pre-training and fine-tuning. 11) **HCRP** (Han et al., 2021), introducing three modules containing Hybrid Prototype Learning, Relation-Prototype Contrastive Learning, and Task Adaptive Focal Loss for the model improvement.

To be noted, **MTB**, **CP** and **MapRE** all employ additional pre-training on BERT with Wikipedia data or contrastive learning to get better contextual representation. Moreover, we respectively use the original BERT and CP as our back-end language models. Therefore, among the 11 baselines mentioned above, **Proto-BERT** and **CP** are our most basic baseline.

5.3.2 NOTA Scenario

We compare our proposed method with three baselines in NOTA scenario: 1) **BERT-PAIR** (Gao et al., 2019b), a method that measures the similarity of sentence pairs. 2) **MNAV**, the Rank 1 method reported on *CodaLab*. 3) **lfc**, the Rank 2 method reported on *CodaLab*.

6 Main Results

6.1 General Scenario

All experimental results are shown in Table 1. CNN-based and BERT-based methods are both contained in the table. *Proto-BERT* represents the method on which our model is based, which means that this is the result of the model without introducing any improvement we propose. This result will also be analyzed and displayed in the ablation study. We apply our proposed method to BERT and CP. For obvious comparisons, the former is shown in the first part of BERT-based models, and the latter is shown in the second part of BERT-based models. The **last two rows** show the increase on the test set compared to the basic models used by our method (i.e., Proto-BERT and CP).

There are several observations. We can observe that, regardless of using BERT or CP, our proposed model (i.e., PRM) outperforms all strong baselines. Particularly, when compared to the base model (i.e., Proto-BERT and CP), PRM achieves significant improvements, as shown in the last two rows of Table 1, further confirming the effectiveness of our innovation in explicitly utilizing relation information to guide prototype representation learning. Besides, the performances gains from the 5-w-1-s, 10-w-1-s setting over the current state-of-the-art model (i.e., HCRP) are larger than that of 5-w-5-s, 10-w-5-s, indicating that PRM is more suitable for the few-shot setting. The possible reason might be that when only one instance is given for each relation class (i.e., 1-shot condition), the *Original Prototype* is the statement representation of the one instance, which is more likely to deviate from the class center. Explicit constraints in PRM have a strong ability to pull this *Original Prototype* closer to the class center, while implicit constraints in HCRP are limited to dealing with such conditions.

6.2 NOTA Scenario

Results are shown in Table 2. **MNAV** and **lfc** are derived from *CodaLab* competition website that are 1st and 2nd methods respectively. The proposed PRM outperforms these methods with both BERT and CP baseline models, which demonstrates the effectiveness of PRM.

7 Analysis

7.1 Ablation Study

In order to analyze the effect of each component in our model (i.e., relations combination way and Prototype Rectification Method/Module), we conducted ablation study experiments on FewRel 1.0 and the results are reported in 3. Since labels of the test set of FewRel are not accessible, ablation experiments are performed on the validation set. There are two parts in Table 3. *Proto-BERT* means the baseline method without any relation information. The second part in the table represents the result obtained after adding a certain module, i.e., relations and PRM. In **w/ relation info.**, instead of PRM, we perform two simple way: 1) **Add**: add relation representations to *Original Prototypes* directly. 2) **Concat**: concatenate relation representations and *Original Prototypes*, then through a linear layer for dimension reduction.

	HCRP	PRM
Para.	110.66M	109.49M
Parameters to be adjusted		
Training	learning rate batch size max iteration	
Loss	λ γ	none

Table 4: Comparison on the model complexity.

Model	5w1s	5w5s	10w1s	10w5s	Aver.
Proto-ADV	42.21	58.71	28.91	44.35	43.55
BERT-PAIR	67.41	78.57	54.89	66.85	66.93
HCRP (BERT)	**76.34**	83.03	**63.77**	72.94	74.02
PRM (BERT)	73.98	**88.38**	62.72	**79.43**	**76.13**

Table 5: Accuracy (%) of few-shot classification on the FewRel 2.0 *domain adaptation* test set, where Aver. denotes the average value of four settings.

We can obtain several observations. First, **Add** achieves better results than Proto-BERT, which proves that the idea of using relations to directly rectify the *Original Prototypes* is indeed effective. However, **Concat** obtains relatively poor results and is even inferior to the original *Proto-BERT*. A possible reason might be **Concat** needs to introduce an extra linear layer to reduce the dimension and thus bring some harmful parameters. Second, **w/ PRM** obtain further improvements on four settings, which demonstrates the effectiveness of PRM.

7.2 Model Complexity and Transferability

As we mentioned in Section 1, our method is simpler and has good transferability compared to the state-of-the-art, i.e., HCRP.

Complexity. The parameters of the two models are shown in Table 4. HCRP utilized three modules to jointly improve the model results, i.e., hybrid features generation, relation-prototype contrastive learning, and task adaptive loss function, whereas PRM only uses the GRM. More details and comparisons can be found in Appendix A, where PRM is compared with different modules of HCRP.

Transferability. We have already demonstrated that PRM still works in the NOTA scenario. In this section, we conduct experiments on domain adaptation settings with FewRel 2.0 and compare the results with HCRP, which is shown in Table 5. It can be seen that PRM is overall better than HCRP, which shows that PRM is also transferable to the domain adaptation setting. The possible reason why PRM is worse than HCRP on 5w1s and 10w1s is that the FewRel 2.0 with domain adaptation setting only provides the name of relations without a specific description, which causes the model to fail to generate a strong relation representation for rectifying the prototypes.

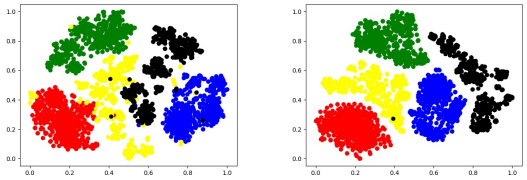

Figure 4: Prototypes Visualization. Left: Original Prototypes; Right: Rectified Prototypes

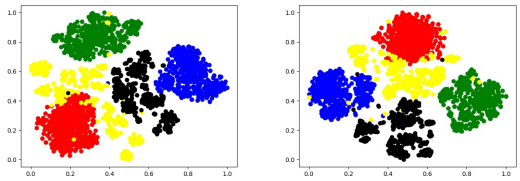

Figure 5: Instances Visualization. Left: Instances in Proto-BERT; Right: Instances in proposed PRM.

7.3 Visualization

In order to further explore how PRM uses relations to rectify the original prototypes, we give the visualization results in Fig. 4, 5 with BERT on 5-way 1-shot of the validation set of FewRel 1.0. Fig. 4 and Fig. 5 show the visualization of prototypes and query instances respectively, where different colors represent different relation classes. From left to right in figures, **Left** means the original prototypes or statements of query instances without any relation information, **Right** means rectified prototypes or statements of query instances with the proposed PRM.

It can be seen that when the relations are not introduced into the model (**Left**), although the prototypes and instances can also be divided into different classes, the intra-class distances are not close enough, and there are multiple error points (i.e., black points). After introducing the relation information (**Right**), we can see that the error points are reduced while the representations of the same class are closer, especially for prototypes in Fig. 4. The observation shows that our proposed method of explicitly introducing relations has a part of the role of contrastive learning and is indeed beneficial to the improvement of the model.

8 Conclusion

In this paper, we proposed a prototype rectification method, PRM, with relations based on prototype framework, where a prototype rectification module is used for obtaining rectified prototypes. We further extended PRM to a none-of-the-above (NOTA) setting in few-shot learning. Extensive experiments demonstrate the effectiveness of the proposed method. We believe that the idea of finding global information to rectify prototypes explicitly with fewer parameters is general and can be extended to other few-shot tasks.

Acknowledgments

This work is supported by Chinese Key-Area Research and Development Program of Guangdong Province (2020B0101350001), NSFC under the project "The Essential Algorithms and Technologies for Standardized Analytics of Clinical Texts" (12026610) and the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. 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. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Yehuda Dar, Vidya Muthukumar, and Richard G Baraniuk. 2021. A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning. *arXiv preprint arXiv:2109.02355*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Bayu Distiawan, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240.

Manqing Dong, Chunguang Pan, and Zhipeng Luo. 2021. 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.

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

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

Victor Garcia and Joan Bruna. 2018. Few-shot learning with graph neural networks. In *6th International Conference on Learning Representations, ICLR 2018*.

Jiale Han, Bo Cheng, and Wei Lu. 2021. Exploring task difficulty for few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2605–2616.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. *arXiv preprint arXiv:2004.03186*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. 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.

Lifeng Jin, Linfeng Song, Yue Zhang, Kun Xu, Weiyun Ma, and Dong Yu. 2020. Relation extraction exploiting full dependency forests. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8034–8041.

Shantanu Kumar. 2017. A survey of deep learning methods for relation extraction. *arXiv preprint arXiv:1705.03645*.

Ziran Li, Ning Ding, Zhiyuan Liu, Haitao Zheng, and Ying Shen. 2019. Chinese relation extraction with multi-grained information and external linguistic knowledge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4377–4386.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from context or names? an empirical study on neural relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672.

Meng Qu, Tianyu Gao, Louis-Pascal Xhonneux, and Jian Tang. 2020. Few-shot relation extraction via bayesian meta-learning on relation graphs. In *International Conference on Machine Learning*, pages 7867–7876. PMLR.

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

Joaquin Vanschoren. 2018. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*.

Ricardo Vilalta and Youssef Drissi. 2002. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95.

Yuxia Wang, Karin Verspoor, and Timothy Baldwin. 2020. Learning from unlabelled data for clinical semantic textual similarity. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 227–233.

Kaijia Yang, Nantao Zheng, Xinyu Dai, Liang He, Shujian Huang, and Jiajun Chen. 2020. Enhance prototypical network with text descriptions for few-shot relation classification. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2273–2276.

Shan Yang, Yongfei Zhang, Guanglin Niu, Qinghua Zhao, and Shiliang Pu. 2021. Entity concept-enhanced few-shot relation extraction. *arXiv preprint arXiv:2106.02401*.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-level matching and aggregation network for few-shot relation classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2872–2881.