A Two-phase Prototypical Network Model for Incremental Few-shot Relation Classification

Haopeng Ren[1,2], **Yi Cai**[1,2]*, **Xiaofeng Chen**[1,2], **Guohua Wang**[1,2], **Qing Li**[3]
[1]School of Software Engineering, South China University of Technology, Guangzhou, China
[2]Key Laboratory of Big Data and Intelligent Robot (South China University of Technology), Ministry of Education
[3]Department of Computing, The Hong Kong Polytechnic University , Hong Kong, China
ycai@scut.edu.cn

Abstract

Relation Classification (RC) plays an important role in natural language processing (NLP). Current conventional supervised and distantly supervised RC models always make a *closed-world* assumption which ignores the emergence of novel relations in an open environment. To incrementally recognize the novel relations, current two solutions (i.e, *re-training* and *lifelong learning*) are designed but suffer from the lack of large-scale labeled data for novel relations. Meanwhile, prototypical network enjoys better performance on both fields of deep supervised learning and few-shot learning. However, it still suffers from the incompatible feature embedding problem when the novel relations come in. Motivated by them, we propose a two-phase prototypical network with prototype attention alignment and triplet loss to dynamically recognize the novel relations with a few support instances meanwhile without catastrophic forgetting. Extensive experiments are conducted to evaluate the effectiveness of our proposed model.

1 Introduction

Relation Classification (RC) is a fundamental task in natural language processing (NLP), aiming to assign semantic relations to the entity pairs mentioned in sentences. Currently, conventional supervised (Zeng et al., 2014; Chen et al., 2020) or distantly supervised (Mintz et al., 2009; Zhang et al., 2019) RC models are widely used and achieve remarkable performance. They are always based on a *closed-world* assumption that the relations expressed in query instances must have appeared in the pre-defined relation set. It is clearly limited in many realistic scenarios, especially in a dynamic or open environment. As shown in Table 1, the query instance **Q** expresses the relation *father* which is out of the pre-defined relation set. However, current supervised RC models ignore the novel relations (i.e., out of the pre-defined relations) and incorrectly classify this query instance into one of the pre-defined relations. To incrementally recognize the novel relations, two kinds of solutions, i.e. *re-training* (Gidaris and Komodakis, 2018) and *lifelong learning* (Wang et al., 2019; Han et al., 2020), are proposed. However, these two solutions still suffer from the lack of large-scale labeled data for novel relations. They are prone to overfitting on novel relations and may even lead to catastrophic forgetting on base ones (i.e., previous pre-defined relations) when given insufficient training data for novel relations (Xiang et al., 2019).

In contrast, it is intuitive that the human can learn new knowledge after being taught just few instances (Snell et al., 2017). Based on this intuition, a series of few-shot RC models (Mishra et al., 2017; Gao et al., 2019a; Soares et al., 2019; Gao et al., 2020) are proposed and can effectively recognize the novel relations with only a few (e.g., 1 or 5) support instances. Nevertheless, current few-shot RC models only focus on learning novel relations and ignore a fact that many common relations (i.e., base relations) are readily available in large datasets. They neglect the existence of large-scale training data for base relations and still learn the base relations in the low-resource setting (i.e., each base relation is given only a few support instances), which cannot fully capture or model the features of base relations. To tackle this limitation of current few-shot RC models, our work in this paper considers a more realistic

*Corresponding author

Proceedings of the 28th International Conference on Computational Linguistics, pages 1618–1629
Barcelona, Spain (Online), December 8-13, 2020

Relation Set	Sentence Sample
capital of	**S1**:[**London**]$_{e1}$ is the capital of the [**U.K**]$_{e2}$
data of birth	**S2**: [**Mark Twain**]$_{e1}$ was born in [**1835**]$_{e2}$
member of	**S3**:[**Newton**]$_{e1}$ served as the president of the [**Royal Society**]$_{e2}$
Query Set	**Q**:[**Gerard**]$_{e1}$ was the father of [**Gottfried**]$_{e2}$

Table 1: A relation classification example in an open environment. For the closed-world assumption based RC models, the set of predefined relations (i.e., *capital of*, *data of birth* and *member of*) is fixed after model training. $[.]_{e1/e2}$ denotes the entity name mentioned in the corresponding sentence.

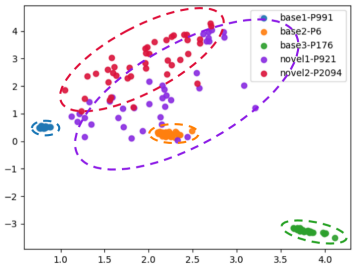


Figure 1: Incompatible Feature Embedding Space. Five relations (3 base and 2 novel relations) with 30 instances are randomly selected from dataset *FewRel* (Gao et al., 2019a) and encoded by prototypical network (Yang et al., 2018). *P991*, *P6*, *P176*, *P921* and *P2094* respectively represent relation *successful candidate*, *head of government*, *manufacturer*, *main subject* and *competition class*.

setting where the relation learning system is able to enjoy both the ability to learn from large-scale data for base relations and the flexibility of few-shot learning for novel ones. Specifically, the RC model not only can learn the base relations from large-scale training data, but also can dynamically recognize the novel relations with only a few support instances. Research on this subject can be named as *incremental few-shot relation classification*.

In both fields of deep supervised learning and few-shot learning, prototypical networks (Snell et al., 2017) obtain better performance on several benchmarks. They conduct classification by learning the distance distribution among relations. However, limited by the *closed-world* assumption, they only focus on the feature embedding learning for the base relations. When the novel relations come in, the feature spatial distributions of novel relations might be distorted and become incompatible with those of base relations. As shown in Figure 1, the base relations are well-distinguished in the feature embedding spaces. Nevertheless, as the novel relations come in, the feature spatial distributions of novel relations are extremely wider than those of base ones and even overlap the spatial distributions of the base relations. It becomes infeasible to conduct classification simultaneously for base and novel relations. To solve this incompatible feature embedding problem, the *prototype attention alignment* (ProtoAtt-Alignment) and *triplet loss function* are designed in our proposed model. They aim to force the prototypical network to narrow down the feature spatial distributions of novel relations and meanwhile to enlarge the distances among different relations in the same embedding space.

In our paper, we propose a two-phase prototypical network model with *ProtoAtt-Alignment* and *triplet loss* for incremental few-shot relation classification. The whole framework is shown in Figure 2. In the **first phase**, a *deep prototypical network* is proposed to learn the feature embedding space of base relations in a supervised learning manner, following Yang et al. (2018). Each base relation is represented as the center (base prototype) of its training instances. To dynamically recognize the novel relations, the *novel prototype generator* is designed to learn the representations for novel relations (novel prototype) with only a few support data. Then, an *incremental prototypical network* with *novel prototype generator* is proposed in the **second phase** and classification is conducted by comparing the distances between query instance and each prototype (i.e., both the base and novel prototypes).

The main contributions of this paper can be summarized as follows: (1) We explore a problem of incremental few-shot relation classification and propose a two-phase prototypical network model to dynamically recognize the novel relations with a few support data meanwhile without catastrophic forgetting. To the best of our knowledge, our work is the first study focusing on *incremental few-shot relation classification*. (2) We design a prototype attention alignment and triplet loss to solve the incompatible feature embedding problem which exists in current prototypical network. (3) Extensive experiments and visualization analysis are conducted on a real-world dataset to evaluate the effectiveness of our model.

2 Related Work

Relation classification (RC) is one of the most important techniques in natural language processing (NLP) and has various applications such as information retrieval (Ercan et al., 2019), question answering (Tong et al., 2019) and dialogue systems (Ma et al., 2019). Currently, conventional deep supervised (Zeng et al., 2014; Gormley et al., 2015) and distantly supervised (Mintz et al., 2009; Jiang et al., 2016; Ye and Ling, 2019a) RC models are widely used and achieve remarkable performance. They are always based on the *closed-world* assumption (Fei and Liu, 2016) that the relation expressed in the query instances must have appeared in the pre-defined relation set. The set of relations which RC models can recognize is fixed after training. However, it is often violated and limited a lot in many realistic scenarios, especially in a dynamic or open environment. Novel relations can emerge dynamically in an open-world scenario.

To dynamically expand the fixed relation set, two solutions can be concluded. Firstly, the base and straightforward method is *re-training* (Gidaris and Komodakis, 2018). Every time the novel relations come in, we need to collect training data for novel relations and then train from scratch on the enhanced training data, aiming to avoid catastrophic forgetting (McCloskey and Cohen, 1989; McClelland et al., 1995). However, the repeating training process is computationally expensive and time-consuming. Recently, two lifelong learning based RC models (Wang et al., 2019; Han et al., 2020) are proposed to alleviate the expensive re-training process. Nevertheless, both the solutions still suffer from the lack of large-scale of training data for novel relations. Without enough training data for novel relations, both the above two solutions risk overfitting on the recognition of novel relations and even suffer from catastrophic forgetting on base relations.

In contrast, humans have the ability to perform even one-shot classification, where only one example of each new category is given. Based on this intuition, a series of few-shot RC models are proposed. They can be classified into two categories: meta-learning based models (Santoro et al., 2016; Ravi and Larochelle, 2016; Mishra et al., 2017) and metric-learning based models (Koch et al., 2015; Snell et al., 2017; Han et al., 2018; Gao et al., 2019a; Fan et al., 2019; Gao et al., 2019b; Soares et al., 2019). However, the few-shot RC models only focus on novel relations learning, but ignore a fact that many common relations are readily available in large datasets. To tackle this problem, we consider a more realistic setting where the relations learning system can not only learn the base relations from the large-scale training data, but also dynamically recognize the novel relations with only a few support examples (termed as *incremental few-shot relation classification*). Currently, several related works (Qi et al., 2018; Gidaris and Komodakis, 2018; Xiang et al., 2019; Ren et al., 2019) are proposed in computer vision field and they concentrate on image classification task. Different from images, the text is more diverse and noisy. It is hard to directly generalize to NLP applications (Gao et al., 2019a). In this paper, we propose a two-phase prototypical network model for incremental few-shot relation classification. Extensive experiments are conducted to evaluate the effectiveness of our proposed model.

3 Model

To address the problem of incremental few-shot relation classification, we propose a two-phase prototypical network model with the *prototype attention alignment* and an auxiliary *triplet loss* (IncreProtoNet). In a dynamic and open environment, novel relations can emerge in test stage. However, current conventional supervised RC models are based on the *closed-world* assumption and neglect the emergence of novel relations. Although current few-shot RC models achieve remarkable performances on recognizing novel relations with a few support instances, they ignore a fact that the common relations (i.e., base relations) are readily available in large datasets. To simultaneously learn the base relations with large training data and dynamically recognize the novel relations with only a few support data, a two-phase IncreProtoNet is proposed, as shown in Figure 2. The first phase, named **deep prototypical network**, is designed to pre-train a base model for base relations in a deep supervised manner. The second phase, named **incremental prototypical network**, is proposed to dynamically recognize the novel relations with only a few support instances meanwhile do not forget the base relations.

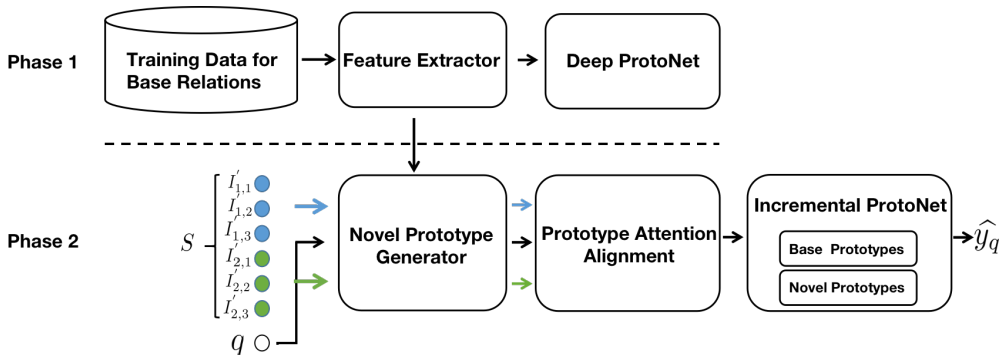


Figure 2: The overview of our proposed framework. Prototypical Network is denoted as *ProtoNet*

3.1 Problem Definitions and Notations

The incremental few-shot relation classification can be defined as a task as follows. We assume there exist a large dataset $D_{train} = \cup_{b=1}^{N_{base}}\{I_{b,i} = (x_{b,i}, h_{b,i}, t_{b,i}, r_b)\}_{i=1}^{K_b}$ of N_{base} base relations, where K_b is the number of training instances of r_b base relation and $I_{b,i}$ means entity pair $(h_{b,i}, t_{b,i})$ mentioned in sentence $x_{b,i}$ which expresses the semantic relation r_b. Using this large training data of base relations, our work aims to effectively learn the base relations and meanwhile to dynamically recognize the novel relations with only a few (e.g., 1 or 5) support instances. Therefore, given a support set S for N_{novel} novel relations, the model can classify the entity pair (h, t) mentioned in query instance q into the most possible relation $r_i \in R_{base} \cup R_{novel}$, where $R_{base} = \{r_b\}_{b=1}^{N_{base}}$ and $R_{novel} = \{r_n^{'}\}_{n=1}^{N_{novel}}$. The support set S can be defined as follows:

$$S = \cup_{n=1}^{N_{novel}}\{I_{n,i}^{'}\}_{i=1}^{K_n^{'}} \tag{1}$$

, where $K_n^{'}$ is the number of support instances of novel relation $r_n^{'}$ and $I_{n,i}^{'}$ is its i-th support instance.

3.2 Feature Extractor

3.2.1 Token Embedding Layer

Given an instance $x = \{w_1, w_2, \ldots, w_L\}$, the token embedding layer aims to transform each input word token w_i into a real-valued vector $\boldsymbol{v}_i \in \mathbb{R}^d$ $(1 \leq i \leq L)$. Following Gao et al. (2019a), the vector $\boldsymbol{v}_i$ consists of two parts: word embedding $\boldsymbol{v}_{w_i} \in \mathbb{R}^{d_w}$ and position embedding $\boldsymbol{v}_{pos_i} \in \mathbb{R}^{2\times d_p}$. We can obtain token representation $\boldsymbol{v}_i$ by concatenating word embedding and position embedding, as follows:

$$\boldsymbol{v}_i = [\boldsymbol{v}_{w_i}; \boldsymbol{v}_{pos_i}]; \boldsymbol{v}_i \in \mathbb{R}^d, d = d_w + 2 \times d_p \tag{2}$$

Finally, each instance can be transformed into an instance matrix $\boldsymbol{S} \in \mathbb{R}^{L\times d}$, where $\boldsymbol{S} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_L\}$ and $\boldsymbol{v}_i \in \mathbb{R}^d$.

3.2.2 Instance Encoder Layer

The instance encoder layer is used to map an instance x into a low-dimensional vector $\boldsymbol{x}$ using a compositional function $f(\boldsymbol{S})$ over the token embedding sequence $\boldsymbol{S}$.

$$\boldsymbol{x} = f_\phi(\boldsymbol{S}) \tag{3}$$

where ϕ is the learnable parameters of compositional function $f(\cdot)$. In our proposed model, we firstly employ convolutional neural networks (CNNs) (Kim, 2014) to capture the local features of instance. The instance matrix $\boldsymbol{S}$ is input into CNNs with d_h filters whose window size is win. It outputs another hidden embedding matrix $\boldsymbol{H} \in \mathbb{R}^{L\times d_h}$. Then, a max-pooling operation is applied over the matrix $\boldsymbol{H}$ to obtain the final instance embedding $\boldsymbol{x} \in \mathbb{R}^{d_h}$.

3.3 Deep Prototypical Network

Prototypical network (Snell et al., 2017; Yang et al., 2018) obtains remarkable performance and enjoys better robustness on several benchmarks. It conducts classification by measuring the distance distribution among relations. In the first phase, following Yang et al. (2018), deep prototypical network with prototype loss is used to train a base model in a deep supervised manner. The goal of this phase is to learn both a good feature extractor and a good base classifier. Given a query instance q from D_{train}, the query instance representation $\boldsymbol{x}_q$ can be obtained by the feature extractor. Then, the probability of query instance q belonging to relation $r_i \in R_{base}$ can be calculated as follows:

$$p_\phi(y = r_i|q) = \frac{\exp(-d(\boldsymbol{x}_q, \boldsymbol{\mu}_i))}{\sum_{j=1}^{|R_{base}|} \exp(-d(\boldsymbol{x}_q, \boldsymbol{\mu}_j))} \tag{4}$$

, where $\boldsymbol{\mu}_i \in \mathbb{R}^{d_h}$ denotes a learnable weight vector of relation $r_i \in R_{base}$ and $d(\cdot, \cdot)$ is the Euclidean distance function for two given vectors. The parameters of the deep prototypical network are learned in this phase and will be freezed after pre-training. Then, we can obtain the prototypes for base relations (base prototypes) by averaging all the available training instance embeddings. They can be denoted as $\boldsymbol{P}_{base} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{N_{base}}\}$.

3.4 Incremental Few-shot Prototypical Network

In order to dynamically recognize the novel relations with only a few support samples, the incremental prototypical network is proposed to learn the features of novel relations and measure their prototypes (novel prototypes). Then, classification can be conducted by measuring the distances between query instance and all the relations' prototypes (i.e., base prototypes and novel prototypes). The second phase mainly consists of two components, including **Novel Prototype Generator** which measures the novel prototypes with a MetaCNN encoder and **Merged Prototypical Network** which merges the base and novel features with a prototype attention alignment.

3.4.1 Novel Prototype Generator

Given a support set $S = \cup_{n=1}^{N_{novel}} \{I'_{n,i}\}_{i=1}^{K'_n}$, each support instance $I'_{n,i}$ is encoded by the Token Embedding Layer in freezed Feature Extractor as a word embedding matrix $\boldsymbol{S}_{n,i}$.

MetaCNN Encoder: As shown in Figure 1, the feature embedding space is distorted a lot when the novel relations come in, which would cause serious classification errors on both base and novel relations. Instead of using the freezed *Feature Extractor*, we build an another MetaCNN encoder to capture the features of novel relations. The network structure is the same as the Instance Encoder Layer using in the base model. Given a word embedding matrix $\boldsymbol{S}_{n,i}$, the MetaCNN encoder can obtain the support instance embedding $\boldsymbol{x}'_{n,i}$.

Feature Averaging Prototype: For each novel relation $r'_n \in R_{novel}$ with K'_n support instances, we can obtain the prototype of novel relation r'_n by $\boldsymbol{p}'_n = \frac{1}{K'_n} \sum_{i=1}^{K'_n} \boldsymbol{x}'_{n,i}$. Then, the novel prototypes can be denoted as $\boldsymbol{P}_{novel} = \{\boldsymbol{p}'_1, \boldsymbol{p}'_2, \ldots, \boldsymbol{p}'_{N_{novel}}\}$.

3.4.2 Merged Prototypical Network with Prototype Attention Alignment

The base and novel prototypes are merged and denoted as $\boldsymbol{P}_{all} = \boldsymbol{P}_{base} \cup \boldsymbol{P}_{novel}$. Given a query instance q, two instance embedding $\boldsymbol{x}_q^{base}$ and $\boldsymbol{x}_q^{novel}$ can be obtained respectively by *Feature Extractor* and *MetaCNN Encoder*. To merge the base and novel features, prototype attention alignment is designed to measure the important degree of base features and novel ones. The merged query instance embedding can be calculated as follows:

$$\boldsymbol{x}_q = \omega_b \boldsymbol{x}_q^{base} + \omega_n \boldsymbol{x}_q^{novel} \tag{5}$$

, where both ω_b and ω_n are scale weight values and $\omega_b + \omega_n = 1.0$. The weight ω_b and ω_n can be measured by the prototype attention alignment as follows:

$$\omega_b = \frac{exp(-d(\boldsymbol{x}_q^{base}, \boldsymbol{v}_{base}))}{exp(-d(\boldsymbol{x}_q^{base}, \boldsymbol{v}_{base})) + exp(-d(\boldsymbol{x}_q^{novel}, \boldsymbol{v}_{novel}))} \text{ and } \omega_n = 1.0 - \omega_b \tag{6}$$

, where $\boldsymbol{v}_{base}$ and $\boldsymbol{v}_{novel}$ respectively denotes the base and novel feature representation. They are respectively calculated as follows:

$$\boldsymbol{v}_{base} = \sum_{i=1}^{N_{base}} \alpha_i \boldsymbol{p}_i \text{ and } \boldsymbol{v}_{novel} = \sum_{i=1}^{N_{novel}} \beta_i \boldsymbol{p}_i^{'} \tag{7}$$

, where α_i denotes the weight value of i-th base prototype and β_i denotes the weight value of i-th novel prototype. The weight value α_i and β_i is calculated as follows:

$$\alpha_i = \frac{exp(-d(\boldsymbol{x}_q^{base}, \boldsymbol{p}_i))}{\sum_{j=1}^{N_{base}} exp(-d(\boldsymbol{x}_q^{base}, \boldsymbol{p}_j))} \text{ and } \beta_i = \frac{exp(-d(\boldsymbol{x}_q^{novel}, \boldsymbol{p}_i^{'}))}{\sum_{j=1}^{N_{novel}} exp(-d(\boldsymbol{x}_q^{novel}, \boldsymbol{p}_j^{'}))} \tag{8}$$

Finally, the probability of query instance q belonging to relation $r \in R_{base} \cup R_{novel}$ can be measured as follows:

$$p_\theta(r|q) = \frac{exp(-d(\boldsymbol{x}_q, \boldsymbol{p}_i^{all}))}{\sum_{j=1}^{N_{base}+N_{novel}} exp(-d(\boldsymbol{x}_q, \boldsymbol{p}_j^{all}))} \tag{9}$$

where $\boldsymbol{p}_i^{all}$ denotes the i-th prototype in $\boldsymbol{P}_{all}$.

3.5 Triplet Loss for IncreProtoNet

The performance of prototypical network highly depends on the spacial distributions of relations in embedding space. To improve the robustness of prototypical network and further solve the incompatible feature embedding problem, the triplet loss function is adopted in our model. Specifically, the target of triplet loss is to force the prototypical network to narrow down the feature spatial distribution of novel relations and meanwhile to enlarge the distances among different relations. Following Fan et al. (2019), the triplet loss function is designed as follows:

$$\pounds_{triplet} = \sum_{i=1}^{M} \sum_{k=1}^{N_{novel}} max(0, \delta + d(g(a_i^k), g(p_i^k)) - d(g(a_i^k), g(n_i^k))) \tag{10}$$

, where M is the total number of training episodes and (a_i^k, p_i^k, n_i^k) is a triplet consists of the anchor, the positive and the negative instances and δ is a hyper-parameter. Note that the anchor is a virtual instance and denotes the novel prototype.

Finally, the loss $\pounds$ in the second phase is a trade-off between the softmax cross-entropy loss $\pounds_{softmax}$ of incremental prototypical network and the triplet loss $\pounds_{triplet}$ by a hyper-parameter λ:

$$\pounds = \pounds_{softmax} + \lambda * \pounds_{triplet} \tag{11}$$

4 Experiment

4.1 Experiment Settings

We conduct experiments[1] on a large-scale public dataset (i.e., *FewRel*) to evaluate the effectiveness of our proposed model. Two kinds of pretrained word embedding methods, namely *Glove* (Pennington et al., 2014) and language model *BERT* (Devlin et al., 2018), can be used to initialize word embeddings in our model and are finetuned during the training stage. The out-of-vocabulary (OOV) words are initialized as an uniform distribution with range $[-0.01, 0.01]$. For the triplet loss, the hyper-parameter δ is set as 5.0 and λ is set as 1.0. The stochastic gradient descent (SGD) optimizer with initial learning rate of 0.01 is used to optimize the model parameters.

In our experiments, we evaluate our proposed model in two incremental few-shot settings (i.e., N_{base} base relations and 5 novel relations with 1-shot or 5-shot learning, where N_{base} is 54 for *FewRel*). In the **first phase**, the deep prototypical network (i.e., base model) is trained in a supervised learning manner and is freezed after training. Specifically, the target of the first phase is to learn the parameters ϕ of

[1]Code is available at `https://github.com/betterAndTogether/IncreProtoNet`

Models	1-shot learning			5-shot learning		
	Base	Novel	Both	Base	Novel	Both
Siamese	49.42 ± 0.06	35.06 ± 0.26	48.20 ± 0.16	60.61 ± 0.04	34.94 ± 0.18	58.44 ± 0.13
Proto	43.20 ± 0.12	39.86 ± 0.26	42.91 ± 0.22	66.74 ± 0.05	57.33 ± 0.15	65.94 ± 0.11
LM-ProtoNet	46.17 ± 0.09	42.20 ± 0.11	45.84 ± 0.08	59.46 ± 0.21	48.68 ± 0.11	58.55 ± 0.17
HATT_Proto	51.58 ± 0.11	45.16 ± 0.18	51.03 ± 0.15	67.77 ± 0.13	61.12 ± 0.09	67.20 ± 0.08
MLMAN	53.40 ± 0.15	45.01 ± 0.11	52.69 ± 0.13	68.40 ± 0.15	55.38 ± 0.08	67.30 ± 0.11
Proto(BERT)	69.01 ± 0.07	52.38 ± 0.20	67.60 ± 0.13	75.59 ± 0.04	62.59 ± 0.17	74.49 ± 0.16
BERT-PAIR	76.03 ± 0.05	58.29 ± 0.13	75.30 ± 0.11	80.01 ± 0.03	64.34 ± 0.14	78.68 ± 0.12
ProtoNet(Increment)	75.63 ± 0.04	18.44 ± 0.02	70.78 ± 0.03	75.07 ± 0.03	47.11 ± 0.04	72.70 ± 0.02
Imprint	62.62 ± 0.13	16.79 ± 0.34	58.73 ± 0.27	67.72 ± 0.09	16.49 ± 0.31	63.38 ± 0.25
LwoF	67.92 ± 0.14	40.87 ± 0.22	65.62 ± 0.15	65.77 ± 0.11	65.23 ± 0.21	65.73 ± 0.15
AttractorNet	66.48 ± 0.19	5.32 ± 0.25	61.29 ± 0.23	68.26 ± 0.22	6.45 ± 0.26	62.78 ± 0.24
IncreProtoNet	70.96 ± 0.21	48.38 ± 0.11	69.36 ± 0.15	72.54 ± 0.16	61.57 ± 0.11	71.54 ± 0.13
BERT-IncreProtoNet	**82.10 ± 0.04**	**60.15 ± 0.11**	**80.65 ± 0.10**	**84.64 ± 0.04**	**65.77 ± 0.09**	**82.26 ± 0.08**

Table 2: Average classification accuracy (%) on dataset *FewRel*. The *Novel* columns report the average 5-way 1-shot or 5-shot classification accuracy of novel relations; the *Base* and *Both* columns respectively report the average classification accuracy of base relations and both type of relations. The above results are calculated by sampling 2000 tasks each with 54 base relations and 5 novel relations. Each relation is randomly sampled 5 query instances.

the feature extractor and the prototypes of base relations $\boldsymbol{P}_{base}$. In the **second phase**, the incremental prototypical network is trained by iteratively sampling few-shot episodes and tries to learn the meta-parameters, following Gidaris and Komodakis (2018).

4.2 Datasets and Data Settings

In our experiments, we use accuracy as the metric. To evaluate the effectiveness of our proposed model, extensive experiments are conducted on a large-scale few-shot RC dataset FewRel (Gao et al., 2019a). The dataset totally contains 80 relations and each relation has 700 instances. To satisfy our experimental settings, we split the dataset into three parts: **training set** which consists of 54 relations (i.e., base relations R_{base}) each with 550 instances; **validation set** which consists of 54 relations (i.e., base relations R_{base}) each with 50 instances and 10 relations (i.e., novel relations in validation stage) each with 700 instances; and **testing set** which consists of 54 relations (i.e., base relations R_{base}) each with 100 instances and 16 relations (i.e., novel relations R_{novel} in testing stage) each with 700 instances. There are no-overlapping instances between training, validation and testing dataset.

4.3 Result Analysis

In our experiments, we compare the performance of our proposed model with two groups of models: several few-shot RC models and four incremental few-shot learning models which designed for CV applications as follows:

1. Few-shot Learning: We select several few-shot RC models (which can be adapted into the incremental few-shot scenario) as baselines. To adapt them into the incremental few-shot setting, we train the few-shot RC models on the training set of base relations. Then, both the base and novel relations are recognized with a few support instances in test stage. They are listed as follows: *Siamese* (Koch et al., 2015), *Proto* (Han et al., 2018), *LM-ProtoNet* (Fan et al., 2019), *HATT_Proto* (Gao et al., 2019a), *MLMAN* (Ye and Ling, 2019b), *Proto(BERT)*, *BERT-PAIR* (Gao et al., 2019b).

2. Incremental Few-shot Learning:

   - **ProtoNet(incremental)** (Snell et al., 2017): prototypical network is adapted to incremental few-shot settings. Each base relation is represented as the average embedding (base prototype) over its all training instances. At test stage, the novel relations are also represented as the average embedding (novel prototypes) over a few support instances. Finally, classification is conducted by comparing the distances between the query instance and each relation prototype.

Models	1-shot learning			5-shot learning		
	Base	Novel	Both	Base	Novel	Both
DeepProtoNet	71.38 ± 0.16	22.91 ± 0.31	67.27 ± 0.28	71.44 ± 0.18	34.10 ± 0.33	68.27 ± 0.24
IncreProtoNet†	**71.81 ± 0.11**	26.46 ± 0.21	67.97 ± 0.17	**73.05 ± 0.09**	50.56 ± 0.17	71.14 ± 0.13
IncreProtoNet‡	69.07 ± 0.22	44.15 ± 0.31	66.88 ± 0.26	71.41 ± 0.12	55.10 ± 0.13	70.03 ± 0.11
IncreProtoNet	70.96 ± 0.21	**48.38 ± 0.11**	**69.36 ± 0.15**	72.54 ± 0.16	**61.57 ± 0.11**	**71.54 ± 0.13**
BERT-DeepProtoNet	**86.17 ± 0.07**	6.58 ± 0.15	79.61 ± 0.10	**86.30 ± 0.06**	10.06 ± 0.13	79.84 ± 0.09
BERT-IncreProtoNet†	81.33 ± 0.16	50.40 ± 0.11	78.71 ± 0.13	76.63 ± 0.11	65.51 ± 0.09	75.77 ± 0.10
BERT-IncreProtoNet‡	82.64 ± 0.11	58.08 ± 0.15	80.29 ± 0.13	82.64 ± 0.09	63.99 ± 0.12	81.07 ± 0.08
BERT-IncreProtoNet	82.10 ± 0.04	**60.15 ± 0.11**	**80.65 ± 0.10**	84.64 ± 0.04	**65.77 ± 0.09**	**82.26 ± 0.08**

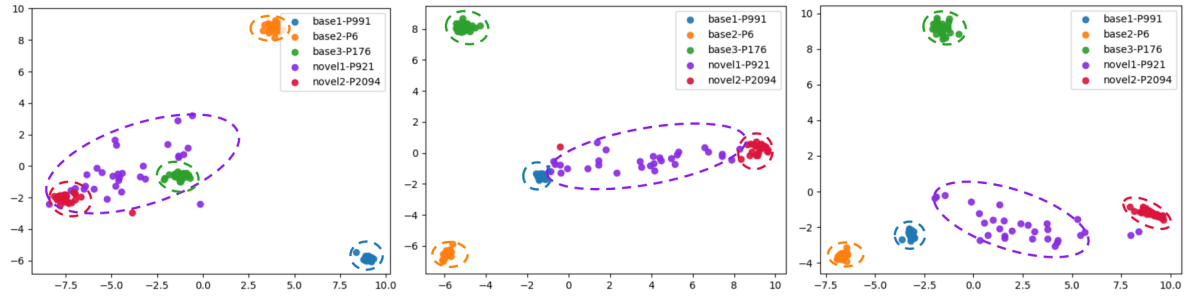
Table 3: Ablation experiments (%) on dataset *FewRel*; *DeepProtoNet* denotes the base model (Yang et al., 2018) which is directly used to recognize both the base and novel relations; † indicates that our model *IncreProtoNet* without *prototype attention alignment*; and ‡ indicates that our model *IncreProtoNet* without *triplet loss function*.

- **Imprint** (Qi et al., 2018): the base classes representations are learned through the supervised pre-training and the novel classes are represented simply by prototypical averaging. Then, classification is conducted over the fully connection layer by concatenating the base and novel classes representations.
- **LwoF** (Gidaris and Komodakis, 2018): Similar to *Imprint*, a two-stage incremental few-shot learning algorithm with a class-wise attention mechanism is designed to learn better classification-weight values for both base and novel classes.
- **AttractorNet** (Ren et al., 2019): Inspired by the attractor networks (Zemel and Mozer, 2001), the attention attractor network model which regularizes the learning of novel classes is designed for incremental few-shot learning on image classification task.

4.3.1 Comparison with Related Models

To demonstrate the effectiveness of our model in the incremental few-shot scenario, we compare our proposed model with two groups of related works (i.e., few-shot RC models and incremental few-shot learning models designed in CV). To adapt the few-shot models to the incremental few-shot settings, both the base and novel relations are recognized in the few-shot learning manner. Recently, a series of works have demonstrated the effectiveness of few-shot learning technique on relation classification task. Nevertheless, they only focus on the novel relation learning and ignore a fact that the common relations (i.e., base relations) have been readily available in large datasets. Specifically, the large-scale training data of base relations is neglected and each base relation is still recognized with only a few support instances. As shown in Table 2, our proposed model achieves higher accuracy by a significant margin on the recognition of base relations. Meanwhile, the novel relations can be also effectively recognized and our model even obtains better performance than current few-shot RC models. Through the comparison with current few-shot RC models, it can demonstrate that our proposed model can not only effectively recognize the base relations, but also dynamically learn the novel relations with only a few support instances.

For the second related works, four incremental few-shot learning models proposed on computer vision field are also implemented and adapted to the relation classification task as the baselines. Different from images, text is more diverse and noise (Gao et al., 2019a). Current incremental few-shot learning models focusing on image classification task are hard to generalize to NLP tasks. From the experimental results shown in Table 2, our proposed model achieves better recognition performance by a significant margin on all experimental settings. Specifically, the model *proto(increment)* encodes the novel relations simply by the pre-trained prototypical network (i.e., base model) and suffers from the incompatible feature embedding problem. Comparing with *proto(increment)*, our proposed model obtains higher accuracy by a large margin on the recognition of novel relations. To some extent, we can conclude that our proposed model can effectively learn compatible feature embedding spaces when the novel relations incrementally come in. The intuitive and specific visualization analysis are given in the final section.



(a) Without *ProtoAtt-Alignment* (b) Without *Triplet Loss* (c) IncreProtoNet

Figure 3: Visualization Analysis; Five relations (i.e., 3 base relations and 2 novel relations) with 30 instances are randomly selected from real-world dataset *FewRel*. They are encoded by our proposed model in three settings and mapped into the 2-dimensional embedding space using PCA algorithm.

4.3.2 Ablation Studies

As shown in Table 3, the ablation experiments on *prototype attention alignment* and *triplet loss* are conducted. The target of the above two components (i.e., ProtoAtt-Alignment and Triplet loss) is to learn the compatible and adaptive embedding spaces when novel relations come in. From the experimental results, both the *ProtoAtt-Alignment* and *Triplet loss* can significantly improve the recognition performance of novel relations and maintain comparable recognition performances of base relations. Especially, the base model (i.e., deep prototypical network) achieves better recognition performance of base relations than our model in some experimental settings. However, it seriously suffers from the incompatible feature embedding problem when novel relations are directly added into the base (initial) embedding space, as shown in Figure 1. Thus, the base model *deepProtoNet* gets the lowest accuracy on the recognition of novel relations in all experimental settings. Through the ablation studies, we can conclude that *prototype attention alignment* and *triplet loss* can effectively force the prototypical network to learn the compatible feature embedding space in the incremental few-shot scenario.

4.3.3 Visualization Analysis

To specifically and intuitively explain the effectiveness of our proposed model, we randomly select 30 instances from the corresponding relations (i.e., 3 base relations and 2 novel relations) in dataset *FewRel* and encode them into the hidden embeddings after the model training. Then, we map them into 2-dimensional points using Principal Component Analysis (PCA) in the same feature embedding space. As shown in Figure 1, current prototypical networks suffer from the incompatible feature embedding problem when the novel relations are directly added into the base embedding space. To solve this problem, *prototype attention alignment* is proposed to learn the compatible or adaptive feature embedding space through aligning and combining the novel and base relations features. Specifically, the feature spatial distribution of both novel and base relations become intra-relation compact and inter-relation separable Comparing with Figure 3(a) and 3(c), it can evaluate the effectiveness of *prototype attention alignment*. The feature spatial distribution of both novel and base relations can be effectively distinguished. What's more, the triplet loss is able to further force the prototypical network to enlarge the distances among relations and shorten the distances within the same relation, as shown in Figure 3(b) and 3(c). It is also beneficial for our proposed model to avoid the catastrophic forgetting on base relations.

5 Conclusion

In this paper, we propose a two-phase prototypical network model for incremental few-shot relation classification. Current conventional supervised RC models are always based on the closed-world assumption that the relations expressed in query instances must have appeared in the pre-defined relations. However, novel relations often emerge in the dynamic or open-world environment. Although current few-shot RC models effectively recognize the novel relations with only a few support instances, they ignore a fact that the common relations (i.e., base relations) are readily available in large datasets. To simultaneously learn

the base relations with large-scale training data and the novel relations with a few support data, an incremental few-shot relations learning model is proposed in our paper. The extensive experimental results and visualization analysis show that our proposed model can effectively recognize the novel relations with a few support data and maintain high recognition accuracy on base relations.

Acknowledgement

This work was supported by the Fundamental Research Funds for the Central Universities, SCUT (No.2017ZD048, D2182480), the National Key Research and Development Program of China, the Science and Technology Programs of Guangzhou (No.201704030076, 201802010027, 201902010046), National Natural Science Foundation of China (62076100) and the Hong Kong Research Grants Council (project no. C1031-18G).

References

Yanping Chen, Kai Wang, Weizhe Yang, Yongbin Qing, Ruizhang Huang, and Ping Chen. 2020. A multi-channel deep neural network for relation extraction. *IEEE Access*, 8:13195–13203.

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

Gonenc Ercan, Shady Elbassuoni, and Katja Hose. 2019. Retrieving textual evidence for knowledge graph facts. In *European Semantic Web Conference*, pages 52–67. Springer.

Miao Fan, Yeqi Bai, Mingming Sun, and Ping Li. 2019. Large margin prototypical network for few-shot relation classification with fine-grained features. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2353–2356. ACM.

Geli Fei and Bing Liu. 2016. Breaking the closed world assumption in text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 506–514.

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 Thirty-Second AAAI Conference on Artificial Intelligence,(AAAI-19), New York, USA*.

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*.

Tianyu Gao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2020. Neural snowball for few-shot relation learning. In *AAAI*, pages 7772–7779.

Spyros Gidaris and Nikos Komodakis. 2018. Dynamic few-shot visual learning without forgetting. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4367–4375.

Matthew R Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. *arXiv preprint arXiv:1505.02419*.

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. *arXiv preprint arXiv:1810.10147*.

Xu Han, Yi Dai, Tianyu Gao, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2020. Continual relation learning via episodic memory activation and reconsolidation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6429–6440, Online, July. Association for Computational Linguistics.

Xiaotian Jiang, Quan Wang, Peng Li, and Bin Wang. 2016. Relation extraction with multi-instance multi-label convolutional neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1471–1480.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. pages 1746–1751.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille.

Mingyu Derek Ma, Kevin Bowden, Jiaqi Wu, Wen Cui, and Marilyn Walker. 2019. Implicit discourse relation identification for open-domain dialogues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 666–672.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. 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: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2017. A simple neural attentive meta-learner. In *ICLR*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Hang Qi, Matthew Brown, and David G Lowe. 2018. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830.

Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning.

Mengye Ren, Renjie Liao, Ethan Fetaya, and Richard Zemel. 2019. Incremental few-shot learning with attention attractor networks. In *Advances in Neural Information Processing Systems*, pages 5276–5286.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087.

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.

Peihao Tong, Qifan Zhang, and Junjie Yao. 2019. Leveraging domain context for question answering over knowledge graph. *Data Science and Engineering*, 4(4):323–335.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Sentence embedding alignment for lifelong relation extraction. *arXiv preprint arXiv:1903.02588*.

Liuyu Xiang, Xiaoming Jin, Guiguang Ding, Jungong Han, and Leida Li. 2019. Incremental few-shot learning for pedestrian attribute recognition. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3912–3918. AAAI Press.

Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. 2018. Robust classification with convolutional prototype learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3474–3482.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019a. Distant supervision relation extraction with intra-bag and inter-bag attentions. 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 2810–2819.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019b. Multi-level matching and aggregation network for few-shot relation classification. *arXiv preprint arXiv:1906.06678*.

Richard S Zemel and Michael C Mozer. 2001. Localist attractor networks. *Neural Computation*, 13(5):1045–1064.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. pages 2335–2344.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. 2019. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. 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 3016–3025.