# Incre-ICAPQ: Iterative Cross Alignment and Prototype Quadruplet Loss for Incremental Few-Shot Relation Classification

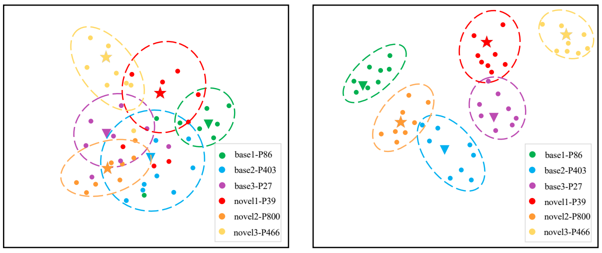
**Anonymous ACL submission**


## Abstract

In incremental few-shot relation classification task, model performance is always limited by incompatibility between base feature embedding space and novel feature embedding space. To tackle the issue, we present a novel method named Incre-ICAPQ with **I**terative **C**ross **A**lignment and **P**rototype **Q**uadruplet loss. Specifically, we incorporate the query instance representation into the encoding of novel prototypes and meanwhile utilize the query-aware prototypes to acquire the query instance representation. To achieve better interaction, we further implement the above dual encoding iteratively. Moreover, prototype quadruplet loss enlarges the distance between different types of prototypes, especially the relative distance between base and novel classes, and makes the distance between query and prototype of the same class as close as possible. Experimental results on two benchmarks demonstrate that Incre-ICAPQ significantly outperforms the state-of-the-art baseline model.


## 1 Introduction

Relation classification (RC), an important sub-task of relation extraction (RE), aims at classifying the relation between two marked entities in a given sentence. For example, the instance "[Newton]$_{e1}$ served as the president of [the Royal Society]$_{e2}$" expresses the relation *member_of* between the two entities *Newton* and *the Royal Society*. Some conventional methods (Zeng et al., 2014; Gormley et al., 2015; Soares et al., 2019) for relation classification adopts supervised training and usually suffer from scarcity of manually annotated data. To alleviate this problem, distant supervision (DS) is adopted to automatically label abundant training instances by heuristically aligning knowledge graphs (KGs) with texts (Mintz et al., 2009). However, the existing DS-based methods fail to deal with the problem of long-tail relations in KGs and still suffer from data deficiency (Han et al., 2018).




(a) IncreProtoNet (b) Incre-ICAPQ

Figure 1: Visualization of the representations of the query instances and prototypes from BERT-IncreProto (a) and our model BERT-Incre-ICAPQ (b). We randomly sample three base relations and three novel relations from real-world dataset FewRel 1.0, each relation with its corresponding prototype (triangles for base relations and stars for novel relations) and eight query instances (points).

To tackle the above long-tail problem, few-shot RC was proposed, which formulates RC in a few-shot learning scenario. This task requires the models trained with base relations to generalize well to novel relations with only few labeled instances. Base relations are those relations containing adequate instances and can be effectively utilized in the training phase to mimic the test phase on novel relations with few samples. Fine-tuning pre-trained models (Bengio, 2012; Gao et al., 2020) is straightforward while suffers from the overfitting problem. Thus, metric based methods (Ravi and Larochelle, 2017; Dong et al., 2020; Geng et al., 2020; Liu et al., 2020b) were proposed to grasp the fast-learning ability from previous experiences and then quickly generalize to new concept. These methods have been experimentally proven to be effective.

Taking a step further, incremental few-shot RC (Ren et al., 2020) considers a more realistic scenario, where the model is required to dynamically recognize the novel relations with a few samples, without reducing the base relation identification capability learned on the large-scale data of base relations. Hence in the test phase, the query set consists of instances of not only base relations but also novel relations, which is more challenging. Several related works (Liu et al., 2020a; Chen and Lee, 2020; Kukleva et al., 2021) have been proposed in the field of computer vision and they focus on image classification task. As for the task of incremental few-shot RC, IncreProtoNet (Ren et al., 2020) is the first work, which proposes a two-phase prototypical network model.

Specifically, IncreProtoNet contains two separate prototypical networks (Snell et al., 2017). One is pre-trained in the first phase to acquire the base prototypes and base feature extractor, and the other obtains the novel prototypes and novel feature encoder with few-shot episode training in the second phase. However, IncreProtoNet suffers from insufficient interaction between the class prototypes and the query instances. Therefore, in the embedding space, novel relations often overlap significantly with base relations, and query representations are scattered, as shown in Figure 1 (a). In addition, the triplet loss used by IncreProtoNet may be affected by noise samples, and its effectiveness decreases on tasks with domain shift. As a result, low accuracy on novel relation recognition has been witnessed.

To alleviate the above problem, we first propose a novel *Iterative Cross Alignment* (ICA) mechanism. Specifically, we build an extra *Cross Alignment* (CA) module to dynamically and interactively encode the novel prototypes and the query instances. On the one hand, the obtaining of novel prototypes is query-aware; namely, the query-related support instances contribute more to the final prototypes. On the other hand, the encoding of query instances is prototype-aware, since the query-related prototypes have more influence on the query representations. Further, we propose to iteratively implement the above CA, namely *Iterative Alignment* (IA), in order to achieve more sufficient interaction and alignment. Besides, *Prototype Quadruplet* (PQ) loss is proposed to enlarge the distance between different types of prototypes, while making the distance between query and prototype of the same class as close as possible.

The contributions of this paper can be summarized below:

- We propose a novel incremental few-shot classification model Incre-ICAPQ with ICA mechanism and PQ loss.
- For the first time, we propose the iterative cross alignment mechinism, which learns the representations of the query instancees and the novel prototypes interactively and iteratively. Besides, a novel prototype quadruplet loss is designed to regulates the feature space distribution.
- Experiments on FewRel 1.0 and 2.0 datasets demonstrate that our method outperforms the state-of-the-art methods by a large margin.

## 2 Task Formulation

In the task of incremental few-shot RC, first we are given a large dataset containing $N_{base}$ base relations: $D_{\text{base}} = \cup_{b=1}^{N_{base}} \{I_{b,i} = (x_{b,i}, h_{b,i}, t_{b,i}, r_b)\}_{i=1}^{K_b}$, in which $K_b$ is the number of instances of relation $r_b$, and $I_{b,i}$ represents its $i$-th instance consisting of the sentence $x_{b,i}$ and the mentioned entity pair $(h_{b,i}, t_{b,i})$. Then we are given a support set $S = \cup_{n=1}^{N_{novel}} \left\{I'_{n,i}\right\}_{i=1}^{K'_n}$ of $N_{novel}$ novel relations, where $K'_n$ is the number of support instances of novel relation $r'_n$ and $I'_{n,i}$ is the $i$-th supporting instance. With $D_{\text{base}}$ and $S$, the task is to recognize the relations of the instances in the query set $Q = \cup_{q=1}^{N_{base}+N_{novel}} \left\{I''_{q,i}\right\}_{i=1}^{K''_q}$, in which $K''_q$ is the number of query instances of relation $r''_q$ and $I''_{q,i}$ is its $i$-th query instance. Therefore, the model is required to dynamically recognize the novel relations based on a few novel support instances while keeping the base relation identification capability learned on the large base dataset.

## 3 Method

In this section, we elaborate on the details of our proposed Incre-ICAPQ method for incremental few-shot RC. First, we give a brief introduction to the IncreProtoNet in Section 3.1. Then, we introduce the overall framework of our model in Section 3.2. Next, we present the proposed ICA mechanism with CA module and IA module in Section 3.3. Moreover, the proposed PQ loss is discussed in Section 3.4.

### 3.1 A Brief Introduction to IncreProtoNet

IncreProtoNet (Ren et al., 2020) is the first work focusing on incremental few-shot RC. The proposed model is a two-phase prototypical network.

In the first phase, a deep prototypical network, consisting of a convolutional neural network based

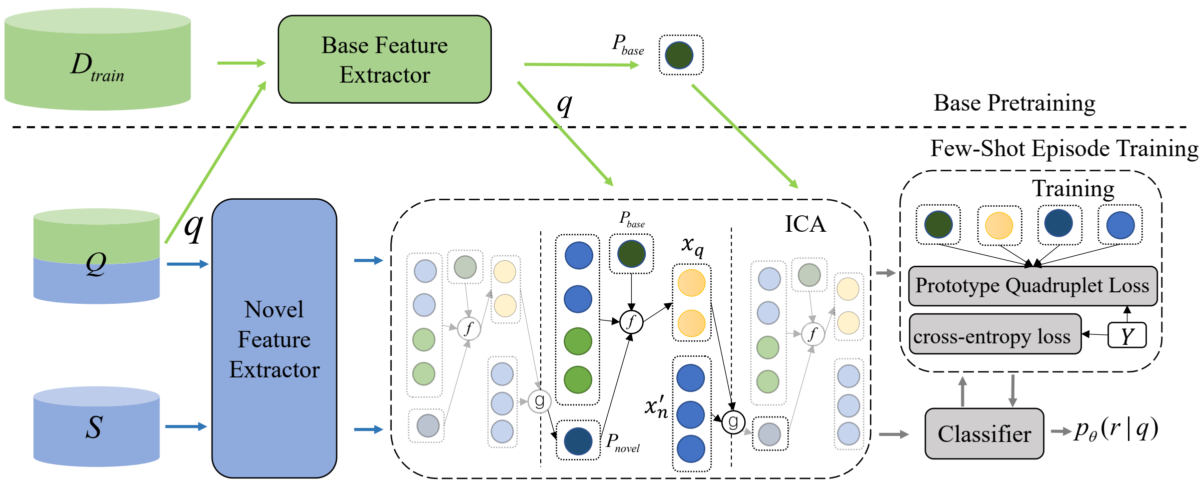



Figure 2: The framework of Incre-ICAPQ. In the dashed box of ICA, ⓕ corresponds to equation (6), which represents the obtaining of query representation; and ⓖ corresponds to equation (8), which represents the update of novel prototypes.

encoder and a prototype based classifier, is pretrained on a large training dataset for base relations in a supervised manner to learn the feature embedding space of base relations. Hence the base prototypes, denoted as $P_{\text{base}} = \{p_1, p_2, \ldots, p_{N_{\text{base}}}\}$, can be obtained by averaging the representations of all the training instances within each base class $b$:

$$p_b = \frac{1}{K_b} \sum_{i=1}^{K_b} x_{b,i}, \tag{1}$$

where $x_{b,i}$ is the embedding of $I_{b,i}$ through the base encoder.

In the second phase, another prototypical network, named incremental few-shot prototypical network, is proposed to learn the feature embedding space of novel relations. The support set is encoded to obtain the novel prototypes $P_{\text{novel}} = \left\{p'_1, p'_2, \ldots, p'_{N_{\text{novel}}}\right\}$ as follows:

$$p'_n = \frac{1}{K'_n} \sum_{i=1}^{K'_n} x'_{n,i}, \tag{2}$$

where $x'_{n,i}$ is the embedding of $I'_{n,i}$ through the novel encoder. For a query instance $q$ from the query set, the representation $x_q$ is calculated as the weighted sum of the $x_q^{base}$ from the base feature embedding space and $x_q^{novel}$ from the novel feature embedding space:

$$x_q = \omega_b x_q^{base} + \omega_n x_q^{novel}, \tag{3}$$

where the weights $\omega_b$ and $\omega_n$ are determined by considering the similarity of the query representation with the base prototypes and novel prototypes, respectively. In short, the query representation calculation can be summarized as:

$$x_q = f(x_q^{base}, x_q^{novel}, P_{base}, P_{novel}), \tag{4}$$

where $f$ is a composite function and represents a series of operations. More details can be found in the original paper (Ren et al., 2020). Lastly, the probability of $q$ belonging to the $i$-th relation $r_i$ can be measured as:

$$p_\theta(r_i \mid q) = \frac{\exp\left(-d\left(\boldsymbol{x}_q, \boldsymbol{p}_i^{all}\right)\right)}{\sum_{j=1}^{N_{\text{base}}+N_{\text{novel}}} \exp\left(-d\left(\boldsymbol{x}_q, \boldsymbol{p}_j^{all}\right)\right)}, \tag{5}$$

where $\boldsymbol{p}_i^{all}$ is the $i$-th prototype in $\boldsymbol{P}_{all} = \{P_{base}, P_{novel}\}$.

Though IncreProtoNet performs well in recognizing instances of base relations, it is still difficult for this model to deal with novel relations. Experimental results in Ren et al. (2020) show that the accuracy for novel relations is much lower than that of base relations, which is unsatisfactory. There are several reasons as follows. First, IncreProtoNet obtains the novel prototypes independent of the query instance, lacking interaction between them. Second, IncreProtoNet ignores the alignment between base relations and novel relations, which is vital in incremental learning scenarios. Third, there is no effective regularization to the feature embedding spaces of base relations and novel relations, which causes discrepancy between them.

### 3.2 Overall Framework of Incre-ICAPQ

To tackle the above issues, we propose the Incre-ICAPQ model on the basis of IncreProtoNet. Like

IncreProtoNet, our model contains two phases including the base pretraining phase and the few-shot episode training phase. Furthermore, we innovatively propose the ICA mechanism and PQ loss, which are demonstrated in the dashed boxes in Figure 2. Next, we illustrate the detailed design.

### 3.3 Iterative Cross Alignment

In the task of incremental few-shot RC, it is important to make an alignment between the base feature embedding space and the novel feature embedding space so as to flexibly encode the query instance and further make correct relation classification. This requires full interaction between base relations and novel relations.

**Cross Alignment**. To this end, the cross alignment (CA) module is designed to encode the novel prototypes and the query instance in an interactive manner. To be specific, we first initialize the novel prototypes $P_{novel}$ and the query instance embedding $x_q$ with equations (2) and (4), respectively. Then, the CA module updates $p'_n \in P_{novel}$, encouraging the model pay more attention to those query-related supporting instances,

$$p'_n = \sum_{i=1}^{K'_n} \gamma_{n,i} x'_{n,i}, \tag{6}$$

where $\gamma_{n,i}$ is defined as:

$$\gamma_{n,i} = \frac{\exp\left(-d\left(x_q, x'_{n,i}\right)\right)}{\sum_{i=1}^{K'_n} \exp\left(-d\left(x_q, x'_{n,i}\right)\right)}, \tag{7}$$

where $d$ is the euclidean distance. In short, the novel prototype embedding process can be summarized as:

$$P_{novel} = g(x_q, \cup_{n=1}^{N_{novel}} \left\{I'_{n,i}\right\}_{i=1}^{K'_n}). \tag{8}$$

Correspondingly, the query instance representation $x_q$ is further updated with equation (4), which requires the model to pay more attention to the query-related base prototypes and novel prototypes. Since most of the query instances belong to base relations, the CA module actually enhances the interaction between instances of base relations and novel relations, achieving better alignment between the two feature embedding spaces.

**Iterative Alignment**. The aligned query representation can help group the different support samples from the same novel class together to optimize the novel prototype. Meanwhile, the optimized novel prototype can further help align query representations from different encoders. Inspired by traditional iterative cross-optimization algorithms, such as the EM (McLachlan and Krishnan, 2007) or *k-means* (Hartigan and Wong, 1979) algorithms, we further propose to carry out the above CA in an iterative way, namely Iterative Alignment (IA). The implementation is straightforward, since we just need to iteratively update $P_{novel}$ and $x_q$ with equations (6) and (4), respectively, until the predefined maximum number of steps is reached. Finally, the refined novel prototypes and query instance representations are obtained. The IA expands CA from single round to multiple rounds, further promoting the interaction and alignment. Algorithm 1 outlines the key steps of our ICA mechanism.

**Algorithm 1** Iterative Cross Alignment

**Input**: Base prototypes $P_{base}$, support set $S$, query instance $q$ and predefined maximum iteration number $N$.
**Parameter**: Base encoder $\Theta_1$ and novel encoder $\Theta_2$.
**Output**: Novel prototypes $P_{novel}$, query instance representation $x_q$ and probability distribution for relation of $q$: $p_\theta(r \mid q)$.

1: Initialize novel prototypes $P_{novel}$ with equation (1).
2: Initialize query instance representation $x_q$ with equation (2).
3: **for** $t = 1 \to N$ **do**
4: Update query representation $x_q^t$:
$x_q^t = f(x_q^{base}, x_q^{novel}, P_{base}, P_{novel}^{t-1})$,
5: Update novel prototypes $P_{novel}^{t+1}$:
$P_{novel}^{t+1} = g(x_q^t, \cup_{n=1}^{N_{novel}} \left\{I'_{n,i}\right\}_{i=1}^{K'_n})$.
6: **end for**
7: **return** $P_{novel}$, $x_q$ and $p_\theta(r \mid q)$.

**Iterative Cross Alignment for Increment Few-Shot Domain Adaptation.** In the real world, especially common in the few-shot scenario, the test (novel classes) domain and training (base classes) domain are often different, so how to improve the abilities of our model to transfer across domains is also very important. Since the test domain usually has no annotations and could differ vastly from the training domain, we first initialize novel class prototypes with average representation of support set instances and the query representations with initialized novel class prototypes. Then the CA module cross-aligns novel support instances and query from different domains. Besides, in the cross-domain scenario, initial query and novel prototypes are more likely to be incompatible; therefore, the ICA mechanism can more significantly improve the representations of the novel prototypes and query from different domains.

### 3.4 Prototype Quadruplet Loss

In our method, there are two feature embedding spaces for base and novel classed separately and the query instance is encoded by the two jointly. Therefore, it is important to measure which embedding space contributes more and further estimate which prototype is the nearest. In addition, the feature spaces of base classes and novel classes should be separated as much as possible when they are embedded into the same space. To this end, we design a novel *Prototype Quadruplet* loss ($\mathcal{L}_{PQ}$), denoted as follows:

$$\begin{aligned}\mathcal{L}_{PQ} = &\sum_{i=1}^{M}\sum_{k=1}^{N_{novel}} \max\left(0, \delta_1 + d_1 - d_2\right) \\ &+ \max\left(0, \delta_2 + d_1 - d_3\right),\end{aligned} \tag{9}$$

where $\delta_1$ and $\delta_2$ are hyper-parameters, $M$ is the total number of training episodes, and three distances $d_1$, $d_2$, $d_3$ are defined as follows:

$$d_1 = d\left(f\left(a_i^k\right), P_{p,i}^k\right), \tag{10}$$

$$d_2 = d\left(f\left(a_i^k\right), P_{n,i}^k\right), \tag{11}$$

$$d_3 = d\left(P_{n,novel,i}^k, P_{n,base,i}^k\right), \tag{12}$$

where $\left(a_i^k, P_{p,i}^k, P_{n,novel,i}^k, P_{n,base,i}^k\right)$ is a quadruplet consisting of the anchor instance, the positive prototype from the same novel class, the first negative prototype from another novel class and the second negative prototype from one of the base classes, $f(\cdot)$ is the feature extractor, and $P_{n,i}^k$ is randomly selected from $P_{n,novel,i}^k$ or $P_{n,base,i}^k$. Different from IncreProtoNet, inspired by the triplet-center loss (He et al., 2018), which can further enhance the discriminative power of the features, we also learn the center representation of each class and then require that the distances between anchors and centers from the same class are smalller than those from different classes. Note that $p^k$, $P_{n,novel,i}^k$, $P_{n,base,i}^k$ are all virtual instances and denote the corresponding prototypes.

In addition, to enhance the abilities of our model to transfer across domains, inspired by the quadruplet loss (Chen et al., 2017) which introduces the absolute distance between the positive and negative sample pairs, we add $d_3$ to better align different domains, which narrows the domain gap and further alleviates the issue of incompatible feature embedding between base classes and novel classes, so as to achieve more effective domain adaptation.

Finally, the joint loss function $\mathcal{L}$ is a trade-off between the cross-entropy loss $\mathcal{L}_{CE}$ and the above $\mathcal{L}_{PQ}$ by a hyper-parameter $\lambda$:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{PQ}. \tag{13}$$

## 4 Experiments

### 4.1 Datasets and Evaluation Metrics

**Datasets.** We carry out extensive experiments on two benchmark datasets. The first one is FewRel 1.0 (Han et al., 2018), which contains 80 relations and provides 700 instances for each relation. We adopt the same split as Ren et al. (2020). To be specific, 54 relations are randomly selected as the base relations each with 550 instances for base pre-training, 50 instances for episode training and 100 instances for testing. 10 other relations each with 700 instances are sampled as the novel relations for the episode training. The rest 16 relations each with 700 instances are used as the novel relations in testing. The other dataset is FewRel 2.0 (Gao et al., 2019b), which is constructed on top of the FewRel 1.0 by adding a new test set in a quite different domain (i.e., medicine), requiring the models to transfer across domains.

**Evaluation Metrics.** To compare our proposed method with the state-of-the-art methods, we adopt the same evaluation metrics as Ren et al. (2020), namely, three kinds of classification accuracy, including classification accuracy for instances of base relations, novel relations, and all relations. Since the number of base relations is much larger than that of novel relations, the classification accuracy for instances of all relations depends largely on that of base relations.

Table 1: Average classification accuracy (%) on the FewRel 1.0 dataset.

| Models | 1-shot learning | | | 5-shot learning | | |
|---|---|---|---|---|---|---|
| | Base | Novel | Both | Base | Novel | Both |
| 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 |
| 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 |
| 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 |
| 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 |
| GloVe-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 |
| GloVe-Incre-ICAPQ | **72.15 ± 0.18** | **54.47 ± 0.04** | **70.42 ± 0.08** | **72.70 ± 0.06** | **71.91 ± 0.10** | **72.63 ± 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 |
| BERT-Incre-ICAPQ | **82.56 ± 0.02** | **63.25 ± 0.09** | **81.50 ± 0.08** | **84.90 ± 0.05** | **69.50 ± 0.06** | **83.64 ± 0.04** |

### 4.2 Implementation Details

To systematically validate the effectiveness of the proposed ICA method, we experiment with two kinds of word embedding initialization methods, namely, GloVe (Pennington et al., 2014) and BERT (Devlin et al., 2019). Besides, the compared methods are all evaluated in both 1-shot and 5-shot learning. The hidden dimension of feature extractor is 230, as well as the prototype dimension. The stochastic gradient descent (SGD) is employed for optimization and the initial learning rate in episode training is set as 0.1, except for BERT as 0.001. For the PQ loss, the two margins $\delta_1$ and $\delta_2$ are set as 5.0 and 10.0 respectively, while the balance weight $\lambda$ is set as 1.

### 4.3 Comparison Methods

First of all, we compare with several few-shot learning models, namely, Proto (Han et al., 2018), HATT Proto (Gao et al., 2019a) and BERT-PAIR (Gao et al., 2019b) and the incremental few-shot learning model ProtoNet (Increment) (Snell et al., 2017). Besides, following (Ren et al., 2020), we compare with Imprint (Qi et al., 2018) and LwoF (Gidaris and Komodakis, 2018) models which are the incremental few-shot learning models in the computer vision field. Finally, we take IncreProtoNet as our baseline, which is the current state of the art.

### 4.4 Main Results

**Our model gains significant improvement in incremental few-show learning tasks.** From Table 1, we can observe that for the FewRel 1.0 dataset, our model achieves the best in both 1-shot and 5-shot tasks. Compared with the best baseline model IncreProtoNet, our model remarkably improves the novel class classification accuracy by 3-10%, while maintaining high accuracy on base class recognition. This shows that the proposed ICA mechanism and PQ loss can greatly promote the models' recognition capabilities for novel classes. We conjecture this is because the ICA mechanism can obtain more effective novel prototypes and better align the query representations from different encoders.

**The more support set instances, the larger the improvement for novel class classification.** As can be seen from Table 1, using either GloVe or BERT as the initial text encoder, the improvement on the 5-shot learning is more significant than that of 1-shot learning for novel class. This is because when there are more support set samples, the ICA mechanism and PQ loss can help separate the base and novel classes, reduce the distance between similar classes, and make the query of novel class and corresponding prototype as close as possible.

### 4.5 Domain Adaptation

To further demonstrate the superiority of our method, we extend the few-shot domain adaptation (few-shot DA) task in FewRel 2.0 (Gao et al., 2019b) to the incremental few-shot domain adaptation (incre-few-shot DA) task in our work. Different from the original incre-few-shot RC, the novel instances in the test set are replaced by new instances from the medical domain. Since the domain of novel instances in the test set is no longer consistent with the training set, the models are required to be able to transfer across domains, which is more challenging.

Table 2 illustrates the comparison results of Incre-ProtoNet and our model, and we have two observations: (1) Huge drops on almost all met-

Table 2: Results (%) of incre-few-shot DA on the FewRel 2.0 dataset.

| Models | 1-shot learning | | | 5-shot learning | | |
|---|---|---|---|---|---|---|
| | Base | Novel | Both | Base | Novel | Both |
| GloVe-IncreProtoNet | 71.37 ± 0.25 | 36.85 ± 0.13 | 68.44 ± 0.18 | 71.71 ± 0.22 | 49.15 ± 0.14 | 69.80 ± 0.17 |
| GloVe-Incre-ICAPQ | **71.39 ±0.11** | **37.03 ± 0.15** | **68.48 ± 0.14** | **73.11 ± 0.15** | **55.58 ± 0.10** | **71.63 ± 0.11** |
| BERT-IncreProtoNet | 86.27 ± 0.06 | 52.68 ± 0.20 | 83.42 ± 0.11 | **87.83 ± 0.05** | 56.70 ± 0.14 | 85.19 ± 0.09 |
| BERT-Incre-ICAPQ | **86.72 ± 0.04** | **52.85 ± 0.16** | **84.58 ± 0.12** | 87.49 ± 0.16 | **65.27± 0.08** | **85.60 ± 0.14** |

Table 3: Ablation Studies. † indicates Incre-ICAPQ without the ICA mechanism; and ‡ indicates Incre-ICAPQ without the PQ loss.

| Models | 1-shot learning | | | 5-shot learning | | |
|---|---|---|---|---|---|---|
| | Base | Novel | Both | Base | Novel | Both |
| GloVe-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 |
| GloVe-Incre-ICAPQ † | 72.03 ± 0.12 | 52.47 ±0.05 | 69.42 ± 0.01 | 72.32 ± 0.04 | 67.36 ± 0.10 | 71.94 ± 0.08 |
| GloVe-Incre-ICAPQ ‡ | 71.15 ±0.03 | 53.97 ± 0.12 | 69.82 ± 0.10 | 71.12 ± 0.06 | 69.14 ± 0.16 | 71.64 ± 0.11 |
| GloVe-Incre-ICAPQ | **72.15 ± 0.18** | **54.47 ± 0.04** | **70.42 ± 0.08** | **72.70 ± 0.06** | **71.91 ± 0.10** | **72.63 ± 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 |
| BERT-Incre-ICAPQ † | 82.20 ± 0.13 | 62.72 ± 0.15 | 80.67 ± 0.08 | 84.04 ± 0.12 | 68.06 ± 0.28 | 82.15 ± 0.10 |
| BERT-Incre-ICAPQ ‡ | 82.15 ± 0.14 | 63.07 ± 0.09 | 80.92 ± 0.13 | **84.98 ± 0.10** | 69.36 ± 0.12 | 83.25 ± 0.15 |
| BERT-Incre-ICAPQ | **82.56 ± 0.02** | **63.25 ± 0.09** | **81.50 ± 0.08** | 84.90 ± 0.05 | **69.50 ± 0.06** | **83.64 ± 0.04** |

rics have been witnessed for both IncreProtNet and our model, which demonstrates the difficulty of incre-few-shot DA. However, the performance of our method deteriorates much slower than that of IncreProtoNet. (2) Our model outperforms IncreProtoNet on all metrics. Especially in 5-shot settings, the accuracy of novel relation recognition is improved by more than 7% in absolute percentage. It indicate that our proposed ICA mechanism provides more accurate, robust and general representations for the relation prototypes and query instances.

### 4.6 Ablation Studies

As shown in Table 3, on the FewRel 1.0 dataset, compared with the baseline IncreProtoNet, our model can get a large improvement with either the ICA mechanism or the PQ loss. Especially for the ICA mechanism, benefited from the full interaction brought by it, better query representation and novel prototype representation greatly improve the model's ability in incremental few-shot learning tasks. Furthermore, these two designs are complementary to each other, and combining them together, we can achieve even larger improvement.

### 4.7 Visualization Analysis

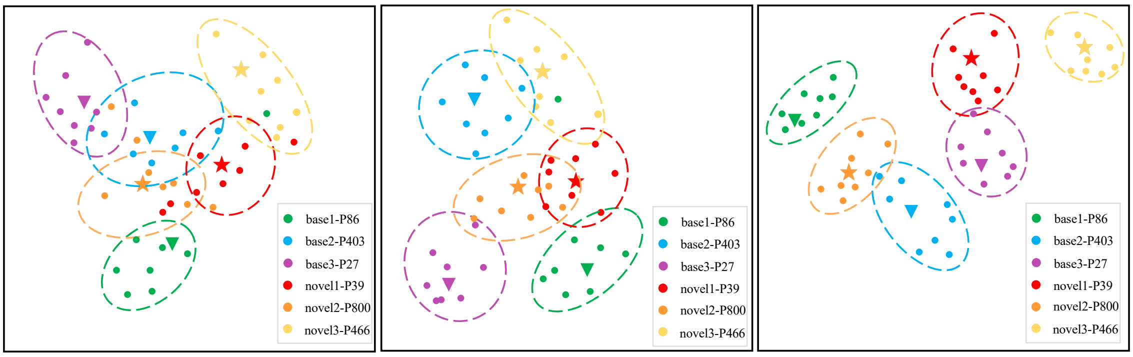
We visualize different types of query representations and prototype representations. As shown in Figure 3, benefited from the ICA mechanism and PQ loss, prototypes of different classes are pushed apart, and the representations of different queries are more accurate and fall close to the corresponding prototype of the same class.

### 4.8 Impact of the Iteration Number of ICA

As shown in Table 4, the ICA mechanism with two (N=2) or three (N=3) iterations achieves better results than the single iteration (N=1). This shows that the ICA mechanism which optimizes query representation and novel prototype representation step by step can effectively improve the accuracy of incremental few-shot learning. In addition, when N is greater than 3, the accuracy of the model decreases. The reason is probably that larger N leads to overfitting of the model. Finally, it can be seen from Table 4 that no matter how many times the model is iteratively aligned, our models are significantly better than the current best baseline IncreProtoNet.

## 5 Related Work

RC is a fundamental task in natural language processing, aiming to recognize the semantic relation between two marked entities in a sentence. With the development of deep learning in recent years, many models based on neural networks have been proposed for this task and achieved great progress.




(a) IncreProto-ICPQ w/o ICA (b) IncreProto-ICPQ w/o PQ loss (c) IncreProto-ICPQ

Figure 3: Visualization of the representations of the query instances and prototypes when BERT-Incre-ICAPQ is equipped without (a) ICA mechanism and without (b) PQ loss.

Table 4: Impact of the iteration number of ICA.

| **Models** | 5-shot learning | | |
|---|---|---|---|
| | Base | Novel | Both |
| GloVe-IncreProtoNet | 72.43 | 61.57 | 71.54 |
| GloVe-Incre-ICAPQ (N=1) | 72.33 | 69.91 | 71.92 |
| GloVe-Incre-ICAPQ (N=2) | 72.55 | 68.91 | 72.45 |
| GloVe-Incre-ICAPQ (N=3) | 72.17 | **71.91** | **72.68** |
| GloVe-Incre-ICAPQ (N=4) | **73.23** | 70.61 | 72.13 |
| BERT-IncreProtoNet | 84.54 | 65.77 | 82.26 |
| BERT-Incre-ICAPQ (N=1) | 84.25 | 67.50 | 82.95 |
| BERT-Incre-ICAPQ (N=2) | 84.36 | **69.50** | 82.24 |
| BERT-Incre-ICAPQ (N=3) | **84.89** | 69.10 | **83.46** |
| BERT-Incre-ICAPQ (N=4) | 84.43 | 68.10 | 82.13 |

For example, Zeng et al. (2014) and dos Santos et al. (2015) utilized convolutional neural networks to capture the global and local semantic information. Later, some attention-based models (Wang et al., 2016; Zhou et al., 2016; Jin et al., 2020) have been proposed to better capture the more useful semantic information. These models are may suffer from the scarcity of high-quality training data. To mitigate the problem, some works (Mintz et al., 2009; Jia et al., 2019; Qin et al., 2018) adopt DS to construct large-scale datasets, while ignore the effect of long-tail relations.

Few-shot RC aims to learn high-quality features with only a small number of training samples. Early works employed the paradigm of pretraining and fine-turning (Bengio, 2012; Donahue et al., 2014; Gao et al., 2020), which aimed to acquire and transfer konwledge from support set containing instances of common relations. Later, metric learning methods (Vinyals et al., 2016; Snell et al., 2017) were proposed to learn different representations across relations. One representative work is prototypical networks (Snell et al., 2017), aiming to learn robust class representations and classify the query set based on the distance to the class prototypes in the feature space. A series of works (Han et al., 2018; Gao et al., 2019a,b) employed prototypical network in few-shot RC and achieved excellent performance.

Incremental learning is a setting where new information is arriving continuously while prior knowledge needs to be maintained. Combining incremental learning with few-shot RC, incremental few-shot RC constitutes a more realistic scenario, where the model is required to leverage the representations of base relations learned from large-scale training dataset meanwhile effectively learn the representations of novel relations from a few support instances. To deal with this task, Ren et al. (2020) proposed a prototypical network based model consisting of two encoders for base relations and novel relations, respectively. In this paper, we argue that the previous work (Ren et al., 2020) is sub-optimal and introduce a preferable solution.

## 6 Conclusion

In this paper, we presented a novel and effective approach with iterative cross alignment and prototype quadruplet loss for the task of incremental few-shot learning. Benefit from the extensive interaction offered by the iterative cross alignment and the feature space regularization brought by the prototype quadruplet loss, our method outperformed the state-of-the-art baseline method significantly, as verified in our extensive experiments. Finally, in our future works, we aim to further improve the performance of our model under the one-shot task setting, as well as accelerate the training process.

## References


Yoshua Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36.

Kuilin Chen and Chi-Guhn Lee. 2020. Incremental few-shot learning via vector quantization in deep embedded space. In *International Conference on Learning Representations*.

Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412.

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.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pages 647–655.

Bowen Dong, Yuan Yao, Ruobing Xie, Tianyu Gao, Xu Han, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2020. Meta-information guided meta-learning for few-shot relation classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1594–1605.

Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. 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)*, pages 626–634.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6251–6256.

Xiaoqing Geng, Xiwen Chen, Kenny Q Zhu, Libin Shen, and Yinggong Zhao. 2020. MICK: A meta-learning framework for few-shot relation classification with small training data. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 415–424.

Spyros Gidaris and Nikos Komodakis. 2018. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE 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. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1774–1784.

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.

John A Hartigan and Manchek A Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Atatistics)*, 28(1):100–108.

Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. 2018. Triplet-center loss for multi-view 3D object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1945–1954.

Wei Jia, Dai Dai, Xinyan Xiao, and Hua Wu. 2019. ARNOR: Attention regularization based noise reduction for distant supervision relation classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1399–1408.

Yanliang Jin, Dijia Wu, and Weisi Guo. 2020. Attention-based LSTM with filter mechanism for entity relation classification. *Symmetry*, 12(10):1729.

Anna Kukleva, Hilde Kuehne, and Bernt Schiele. 2021. Generalized and incremental few-shot learning by explicit learning and calibration without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9020–9029.

Qing Liu, Orchid Majumder, Alessandro Achille, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. 2020a. Incremental few-shot meta-learning via indirect discriminant alignment. In *European Conference on Computer Vision*, pages 685–701. Springer.

Xiaoqian Liu, Fengyu Zhou, Jin Liu, and Lianjie Jiang. 2020b. Meta-learning based prototype-relation network for few-shot classification. *Neurocomputing*, 383:224–234.

Geoffrey J McLachlan and Thriyambakam Krishnan. 2007. *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

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*, pages 1003–1011.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 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.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2137–2147.

Sachin Ravi and Hugo Larochelle. 2017. Optimization as a model for few-shot learning. In *Proceedings of the 5th International Conference on Learning Representations*, pages 224–234.

Haopeng Ren, Yi Cai, Xiaofeng Chen, Guohua Wang, and Qing Li. 2020. A two-phase prototypical network model for incremental few-shot relation classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1618–1629.

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

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.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29:3630–3638.

Linlin Wang, Zhu Cao, Gerard De Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212.