RAPS: A Novel Few-Shot Relation Extraction Pipeline with Query-Information Guided Attention and Adaptive Prototype Fusion

Anonymous ACL submission

Abstract

Few-shot relation extraction (FSRE) aims at recognizing unseen relations by learning with limited annotated instances. Prototypical network has been proven to be simple yet effective method for the FSRE task. Some existing prototypical network-based methods start from implicitly introducing relation information to assist prototype representation learning, which may bring useless and even harmful parameters. Besides, these approaches construct relation prototypes simply by averaging representations of support instances in each relation class, which might not be efficient since the contribution of the support instances are generally heterogeneous. To solve the aforementioned issues, we propose a novel pipeline for the FSRE task based on que**R**y-information guided **A**ttention and adaptive **P**rototype fu**S**ion, **RAPS** for short. Specifically, we first derive a relation prototype by a query-information guided attention module, which exploits rich interactive information between support instances and query instances, in order to obtain more accurate initial prototype representations. Then, we elaborately combine the derived initial prototype with the relation information by an adaptive prototype fusion mechanism to obtain an integrated prototype for both train and prediction. Experiments conducted on two well-known benchmark datasets show a significant improvement of RAPS against previous state-of-the-art methods.

1 Introduction

Relation extraction (RE) is an important part of information extraction in the field of natural language processing (Bach and Badaskar, 2007). It aims at extracting and classifying the relation between two entities contained in a given text and can be applied in other advanced tasks (Li et al., 2022; Hu et al., 2021), such as knowledge graph (Zhao et al., 2020), machine reading comprehension (Ding et al., 2019; Dua et al., 2020), and question answering (Karpukhin et al., 2020; Zhang et al., 2021). Currently, most researches on RE start from deep learning methods, but these methods rely on large-scale and high-quality annotated datasets. In many real-world applications, it is not possible to collect sufficient instances for model training, greatly limiting their applications. In order to solve the problem of data scarcity, Few-shot relation extraction (FSRE) task has been widely studied in recent years. In the FSRE task, a model is first trained on a large-scale annotated data with known relation types, and then quickly adapts to a small amount of data with new relation types.

Support Set
Class 1 mother:
Instance 1 Jinnah and his wife [Rattanbai Petit]$_{e_h}$ had separated soon after their daughter, [Dina Wadia]$_{e_t}$ was born.
Instance 2 She married (and murdered) [Polyctor]$_{e_h}$, son of Aegyptus and [Caliadne]$_{e_t}$. Apollodorus.
Class 2 follows: ...
Class 3 crosses: ...
Query Instance
Dylan and [Caitlin]$_{e_h}$ brought up their three children, [Aeronwy]$_{e_t}$, Llewellyn and Colm.

Table 1: An example of 3-way 2-shot relation classification scenario from the FewRel validation set. Head entity and tail entity are denoted by e_h and e_t, respectively. The query instance is of Class 1: mother. The support instances of Classes 2 and 3 are omitted.

Recently, many approaches have been proposed for addressing FSRE problems (Qu et al., 2020; Peng et al., 2020; Wang et al., 2020; Yang et al., 2020; Han et al., 2021). One of the most popular algorithms is Prototypical Network (Snell et al., 2017), which is a metric based meta-learning framework. The main idea of the prototypical network is that each relation class has a prototype, and the prototype can be learned in the embedding space with given instances (generally average the embedding of the instances in each relation class). All query instances are classified via the nearest neighbor rule.

Two main lines of research tracks have been adopted to improve the FSRE performance. The first line is to integrate the relation information (i.e., relation labels or descriptions) into the model as the external knowledge to assist prototype representation learning. Yang et al. (2020) proposed TD-Proto model, which is an enhanced prototypical network with both relation and entity descriptions. Wang et al. (2020) proposed CTEG model that learns to decouple relations by adding two types of external information. The second line starts from model structure or training strategy to obtain good prototypes by learning intra-class similarity and inter-class dissimilarity. Qu et al. (2020) introduced a global relation graph into a Bayesian meta-learning framework, which makes the resulting model can be better generalized across different relations. Peng et al. (2020) proposed a contrastive pre-training framework for RE to enhance the ability to grasp entity types and extract relational facts from contexts. Han et al. (2021) introduced a novel supervised contrastive learning method that obtains better prototypes by combining prototypes, relation labels and descriptions to support model training, and designed a task adaptive focal loss to improve the performance on hard FSRE task.

However, there are at least two limitations in the above existing works. First, these prototypical-network based methods tend to construct class prototypes simply by averaging representations of support instances of each class, which ignores the informative interaction between support instances and query instances. Second, in order to learn better representations, these works usually adopt complicated designs or networks, such as graphs (Qu et al., 2020), hybrid features (Han et al., 2021), contrastive learning (Wang et al., 2020; Han et al., 2021), or elaborate attention networks (Yang et al., 2021), which may bring too many useless or even harmful parameters.

To address the aforementioned issues, we propose a novel pipeline for the FSRE task based on que**R**y-information guided **A**ttention and adaptive **P**rototype fu**S**ion, **RAPS** for short. Concretely, RAPS exploits rich interactive information between support instances and query instances by a query-information guided attention module to obtain more accurate relation prototype representations. Furthermore, it elaborately combines the derived relation prototype with the relation information by an adaptive prototype fusion mechanism, which provides more degrees of freedom to learn from data by properly weighting the relation prototype and relation information. In this way, the resulting model gains diverse and discriminative prototype representations without introducing too much parameters or computational-demanding modules. Experiments conducted on FewRel 1.0 (Han et al., 2018) and FewRel 2.0 (Gao et al., 2019b) benchmarks show that our model significantly outperforms the previous baselines. Ablation and case studies demonstrate the effectiveness of the proposed modules.

The contributions of this paper are summarized as follows:

- We exploit the rich interactive information between the support set and the query set by the proposed query-information guided attention module to get more accurate prototype.
- We present a novel prototype attention fusion mechanism to further combine the useful information from the relation prototypes and the relation information.
- Qualitative and quantitative experiments on FewRel benchmark demonstrate the effectiveness of the proposed RAPS model.

2 Task Definition

We follow the typical N-way K-shot FSRE task setting, involving a support set $\mathcal{S}$ and a query set $\mathcal{Q}$. The support set $\mathcal{S}$ includes N novel classes (relations), each with K labeled instances. The query set $\mathcal{Q}$ contains the same N classes as $\mathcal{S}$. The FSRE task aims to predict the relation of instances in $\mathcal{Q}$. In addition, an auxiliary dataset $\mathcal{D}_{\text{base}}$ is provided. $\mathcal{D}_{\text{base}}$ contains abundant base classes, each of which has a large number of labeled examples. Note that the relations in $\mathcal{D}_{\text{base}}$ and $\mathcal{S}$ are disjoint. A learner designed for the FSRE task aims to acquire knowledge from the classes in $\mathcal{D}_{\text{base}}$ and make fast adaptation on novel classes in $\mathcal{S}$. Specifically, in each training iteration, N different classes are randomly selected from $\mathcal{D}_{\text{base}}$ to form the support set $\mathcal{S} = \left\{s_k^i; 1 \leq i \leq N, 1 \leq k \leq K\right\}$. Meanwhile, $|\mathcal{Q}|$ instances are sampled from the remaining data of the same N classes to form the query set $\mathcal{Q} = \{q_j; 1 \leq j \leq |\mathcal{Q}|\}$. Each instance in $\mathcal{D}_{\text{base}}$ can be represented as a triple (s, e, r), where s is a sentence of length T, $e = (e_1, e_2)$ are head and tail entities and $r \in \mathcal{R}$ is the semantic relation between

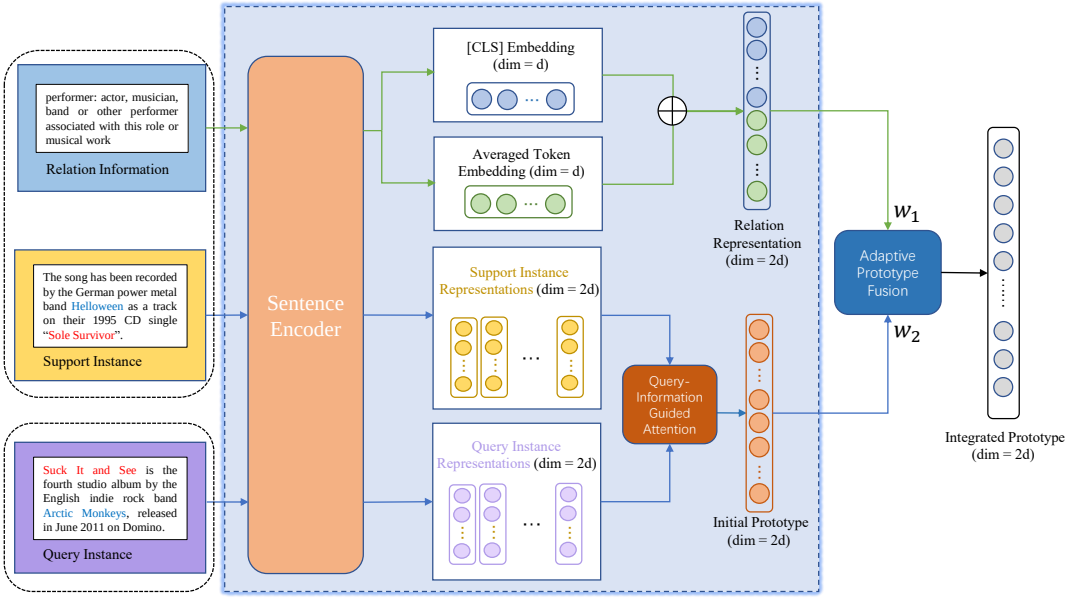

Figure 1: The pipeline of our proposed RAPS model. $\bigoplus$ represents direct concatenation.

e_1 and e_2 conveyed by s, where $\mathcal{R} = \{r_1, \ldots, r_N\}$ is the set of all candidate relation classes. Table 1 is an example of a 3-way 2-shot FSRE task.

3 Methodology

This section provides the details of RAPS. Figure 1 shows the overall model structure. The inputs of this model contain two fractions. 1) the N-way K-shot tasks sampled from $\mathcal{D}_{\text{base}}$, where each task consists of a support set and a query set. 2) the names and descriptions of these N relation classes. We obtain the integrated relation prototypes with abundant information by the following four steps. First, we encode the sentences and relation information into their embeddings by a shared encoder (Sec 3.1). Second, we concatenate two aspects of the relation embeddings to obtain the same dimension as instance representations (Sec 3.2). Third, we calculate the initial prototype of each relation class by the query-information guided attention module (Sec 3.3). Finally, we integrate relation representations into the initial prototypes by an adaptive prototype fusion mechanism to obtain integrated relation prototypes (Sec 3.4).

3.1 Sentence Encoder

We adopt BERT (Devlin et al., 2019) as an encoder to map the instances into a low-dimensional vector space and better capture the semantic information of the support set $\mathcal{S}$ and the query set $\mathcal{Q}$. For instances in $\mathcal{S}$ and $\mathcal{Q}$, the instance representations are obtained by concatenating the hidden states corresponding to start tokens of two entity mentions following Soares et al. (2019): $h = \text{concat}(h_1, h_2) \in \mathbb{R}^{2d}$, where $h_i \in \mathbb{R}^d, i = 1, 2$, and d is the hidden size of BERT output.

3.2 Relation Representation

For each relation, we combine its name and description, and feed the "name-description" sequence into the BERT encoder. We treat the embeddings of the *[CLS]* tokens, $\{\mathbf{r}_i^1, 1 \le i \le N\}$, and the average embeddings of all tokens, $\{\mathbf{r}_i^2, 1 \le i \le N\}$, as two components of the relation representation. Then, the representation of each relation is derived by direct concatenation of $\mathbf{r}_i^1$ and $\mathbf{r}_i^2$, i.e.,

$$\mathbf{r}_i = \text{concat}(\mathbf{r}_i^1, \mathbf{r}_i^2) \in \mathbb{R}^{2d}.$$

3.3 Query-Information Guided Attention Module

Traditionally, the naïve prototype for relation i is computed simply by averaging the embeddings of the K instances under relation i in $\mathcal{S}$ as

$$\mathbf{p}_i = \frac{1}{K} \sum_{k=1}^{K} \mathbf{s}_k^i, \tag{1}$$

where $\mathbf{s}_k^i$ is the embedding of the kth support instance of relation i. However, simply averaging all support instances as the relation prototype will lose abundant interaction between the support and query instances. In other words, different relation

prototypes will be computed based on the semantic relevance between the support and query instances.

In order to generate more informative prototypes, we propose a query-information guided attention module. This specific module aims to utilize the semantic relevance between the support instances and the query instances to help compute the customized relation prototypes. Specifically, the prototype can be represented as

$$\mathbf{p}_i = \sum_{k=1}^{K} \alpha_k^i \mathbf{s}_k^i, \tag{2}$$

where $\mathbf{p}_i$ is the attention-based prototype of relation i, $\mathbf{s}_k^i$ is defined below (1), α_k^i is a weight indicating the semantic relevance between the kth support instance of relation i and the query instances. The weight α_k^i is computed by

$$\alpha_k^i = \frac{\exp\left(-\frac{1}{|\mathcal{Q}|}\sum_{j=1}^{|\mathcal{Q}|} d(\mathbf{s}_k^i, \mathbf{q}_j)\right)}{\sum_{k=1}^{K}\exp\left(-\frac{1}{|\mathcal{Q}|}\sum_{j=1}^{|\mathcal{Q}|} d(\mathbf{s}_k^i, \mathbf{q}_j)\right)}, \tag{3}$$

where $\mathbf{q}_j$ is the embedding of the jth query instance, $|\mathcal{Q}|$ is the total number of instances in $\mathcal{Q}$, and $d(\cdot,\cdot)$ is the Euclidean distance. We call the prototype obtained in this way *Initial Prototype*. When one calculates the prototype for each relation class, taking the semantic relevance between the support instances and the query instances into consideration will improve data utilization efficiency, which is more reasonable than previous works. The benefit of the initial prototype for subsequent classification will be illustrated in Subsection 4.4.

3.4 Adaptive Prototype Fusion

Inspired by the adaptively spatial feature fusion approach proposed by Liu et al. (2019), we design a novel adaptive prototype fusion mechanism to obtain the *integrated prototype*. Specifically, the integrated prototype $\mathbf{p}_i^w$ is an weighted average of $\mathbf{p}_i$ and $\mathbf{r}_i$:

$$\mathbf{p}_i^w = w_1\mathbf{p}_i + w_2\mathbf{r}_i, \tag{4}$$

where $w_1, w_2 \in \mathbb{R}$ are two learnable scalar weights, which depend on the representation of category information to ensure that the calculated integrated prototype $\mathbf{p}_i^w$ is compatible and not redundant. Unlike Liu et al. (2019) forces the sum of the weights equal to 1, our adaptive prototype fusion has no constraint on w_1 and w_2. This provides more degrees of freedom to learn from data and make trade-off between $\mathbf{p}_i$ and $\mathbf{r}_i$. The benefits of mechanism will be demonstrated in Subsection 4.5.

3.5 Training Objective

With the representation of query and integrated prototypes of N relations, our model uses the vector dot product way to compute the probability of the outcome given the query instance representation $\mathbf{q}_j$ as follows:

$$P(y=i|\mathbf{q}_j) = \frac{\exp(\mathbf{q}_j \cdot \mathbf{p}_i^w)}{\sum_{n=1}^{N}\exp(\mathbf{q}_j \cdot \mathbf{p}_n^w)}.$$

The training loss $\mathcal{L}$ is defined as regular cross entropy loss as follows:

$$\mathcal{L} = -\sum_{j=1}^{|\mathcal{Q}|}\log\frac{\exp(\mathbf{q}_j \cdot \mathbf{p}_i^w)}{\sum_{n=1}^{N}\exp(\mathbf{q}_j \cdot \mathbf{p}_n^w)}.$$

In the prediction stage, query $\mathbf{q}_j$ is assigned to relation i with the highest probability:

$$\text{label} = \arg\max_{1\le i\le N} P(y=i|\mathbf{q}_j).$$

4 Experiments

4.1 Datasets and Baselines

Datasets We use FewRel 1.0 (Han et al., 2018) to evaluate our model. FewRel 1.0 is a large-scale FSRE dataset, which contains 100 relations with 700 instances each relation. We follow the official split to use 64, 16 and 20 relations for training, validation, and testing.

In order to study the domain transferability of RAPS, we also evaluate our model on FewRel 2.0 (Gao et al., 2019b). The training set of FewRel 2.0 is the same as FewRel 1.0, while the validation set of FewRel 2.0 is collected from a biomedical domain database called PubMed [1], which contains 10 relations with 100 instances each relation.

Baselines We compare our model with the following baselines. 1) **Proto-CNN** (Snell et al., 2017), prototypical networks with CNN encoder. 2) **Proto-HATT** (Gao et al., 2019a), a hybrid attention applied on prototypical networks to focus on the crucial instances and features. 3) **MLMAN** (Ye and Ling, 2019), a multi-level matching and aggregation prototypical network. 4) **Proto-BERT** (Snell et al., 2017), prototypical networks with BERT encoder. 5) **MAML** (Finn et al., 2017), a model-agnostic meta-learning algorithm. 6) **GNN** (Satorras and Estrach, 2018), a meta-learning approach using graph neural networks. 7) **BERT-PAIR** (Gao

[1] https://www.ncbi.nlm.nih.gov/pubmed/

Encoder	Model	5-way-1-shot	5-way-5-shot	10-way-1-shot	10-way-5-shot
CNN	Proto-CNN♣ (Snell et al., 2017)	72.65 / 74.52	86.15 / 88.40	60.13 / 62.38	76.20 / 80.45
	Proto-HATT (Gao et al., 2019a)	75.01 / — —	87.09 / 90.12	62.48 / — —	77.50 / 83.05
	MLMAN (Ye and Ling, 2019)	79.01 / 82.98	88.86 / 92.66	67.37 / 75.59	80.07 / 87.29
BERT	Proto-BERT* (Snell et al., 2017)	82.92 / 80.68	91.32 / 89.60	73.24 / 71.48	83.68 / 82.89
	MAML* (Finn et al., 2017)	82.93 / 89.70	86.21 / 93.55	73.20 / 83.17	76.06 / 88.51
	GNN* (Satorras and Estrach, 2018)	— — / 75.66	— — / 89.06	— — / 70.08	— — / 76.93
	BERT-PAIR♣ (Gao et al., 2019b)	85.66 / 88.32	89.48 / 93.22	76.84 / 80.63	81.76 / 87.02
	REGRAB (Qu et al., 2020)	87.95 / 90.30	92.54 / 94.25	80.26 / 84.09	86.72 / 89.93
	TD-Proto (Yang et al., 2020)	— — / 84.76	— — / 92.38	— — / 74.32	— — / 85.92
	CTEG (Wang et al., 2020)	84.72 / 88.11	92.52 / 95.25	76.01 / 81.29	84.89 / 91.33
	ConceptFERE (Yang et al., 2021)	— — / 89.21	— — / 90.34	— — / 75.72	— — / 81.82
	HCRP (Han et al., 2021)	90.90 / 93.76	93.22 / 95.66	84.11 / 89.95	87.79 / 92.10
	DRK (Wang et al., 2022)	— — / 89.94	— — / 92.42	— — / 81.94	— — / 85.23
	RAPS	**92.26 / 94.93**	**94.08 / 96.92**	**87.23 / 90.65**	**89.87 / 93.72**
	MTB (Soares et al., 2019)	— — / 91.10	— — / 95.40	— — / 84.30	— — / 91.80
	CP (Peng et al., 2020)	— — / 95.10	— — / 97.10	— — / 91.20	— — / 94.70
	MapRE (Dong et al., 2021)	— — / 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**
	RAPS (CP)	**96.28 / 97.39**	**97.74 / 98.00**	**93.86 / 95.21**	**95.39** / 96.32

Table 2: Accuracy (%) of few-shot classification on the FewRel 1.0 validation / test set. ♣The results of Proto-CNN and BERT-PAIR are from FewRel public leaderboard (https://thunlp.github.io/fewrel.html), *the results of Proto-BERT and MAML are reported in Qu et al. (2020).

et al., 2019b), a method that measures the similarity of sentence pairs. 8) **REGRAB** (Qu et al., 2020), a Bayesian meta learning method with an external global relation graph. 9) **TD-Proto** (Yang et al., 2020), a model learning the importance distribution of generic content words by a memory network. 10) **CTEG** (Wang et al., 2020), a model using dependency trees to learn to decouple high co-occurrence relations, where two external information are added. 11) **ConceptFERE** (Yang et al., 2021), a model introducing the inherent concepts of entities to provide cludes for relation prediction. 12) **HCRP** (Han et al., 2021), a model introducing Hybrid Prototype Learning, Relation-Prototype Contrastive Learining, and Taks Adaptive Focal Loss for the model improvement. 13) **DRK** (Wang et al., 2022), introducing a logic rule to constrain the inference process, thereby avoiding the adverse effect of shallow text features. Moreover, we compare our model with three pretrained RE methods: 13) **MTB** (Soares et al., 2019), a model pretrained by matching the blank stategy on top of an existing BERT model. 14) **CP** (Peng et al., 2020), an entity masked contrastive pretraining framework for RE while utilizing prototypical networks for fine-tuning on FSRE. 15) **MapRE** (Dong et al., 2021), a framework considering both label-agnostic and label-aware semantic mapping information in pre-training and fine-tuning.

Note that MTB (Soares et al., 2019) employs $\text{BERT}_{\text{LARGE}}$ as backbone, and CP (Peng et al., 2020) and MapRE (Dong et al., 2021) all employ additional pre-training on BERT with Wikipedia data or contrastive learning to obtain a better contextual representation. This is the reason why we do not compare RAPS with MTB, CP, and MapRE directly.

4.2 Training and Evaluation

Training The approach is implemented with PyTorch (Paszke et al., 2019) and trained on 1 NVIDIA Geforce RTX 3090 GPU. We use uncased $\text{BERT}_{\text{BASE}}$ and CP (Wang et al., 2020) as the sentence encoder respectively, and set the total training iteration number as 30,000, validation iteration number as 1,000, batch size as 4, learning rate as 1×10^{-5} and 5×10^{-6} for BERT and CP, respectively. The AdamW optimizer (Loshchilov and Hutter, 2019) is applied to minimize the training loss.

Evaluation We evaluate our model on FewRel 1.0 and FewRel 2.0 under multiple N-way K-shot meta tasks. We select N to be 5 and 10, K to be 1 and 5 to form four test scenarios according

Model	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
Proto-CNN*	35.09	49.37	22.98	35.22
Proto-BERT*	40.12	51.50	26.45	36.93
BERT-PAIR*	67.41	78.57	54.89	66.85
Proto-CNN-ADV*	42.21	58.71	28.91	44.35
Proto-BERT-ADV*	41.90	54.74	27.36	37.40
DaFeC+BERT-PAIR♣	61.20	76.99	47.63	64.79
Cluster-ccnet†	67.70	84.30	52.90	74.10
HCRP♠	**76.34**	83.03	**63.77**	72.94
RAPS	74.99	**87.85**	60.29	**80.10**
RAPS (CP)	**80.61**	**89.59**	**67.51**	**82.52**

Table 3: Accuracy (%) of few-shot classification on the FewRel 2.0 domain adaptation test set. *The results are quoted from FewRel leaderboard (https://thunlp.github.io/fewrel.html), ♣the results are quoted from Cong et al. (2021), ♠the results are quoted from Han et al. (2021), †the results are quoted from Bansal et al. (2021).

to Gao et al. (2019a). For FewRel 1.0, we report the accuracy on the validation and test sets. For FewRel 2.0, we report the accuracy on the test set. Since the labels of two test sets are not publicly available, we submit the prediction file of our best model to the CodaLab platform [2] to obtain the prediction results on the test sets.

4.3 Overall Evaluation Results

Table 2 presents the experimental results on the FewRel 1.0 validation set and test set. As shown in the upper part of Table 2, our method evidently outperforms the strong baseline models. To be specific, RAPS achieves an average of 1.19 points improvement in terms of accuracy on the test sets of four meta tasks, compared to the second best method (HCRP), demonstrating its superior generalization ability. In addition, we evaluate RAPS based on CP (Peng et al., 2020), whose BERT encoder is initialized with the pre-trained parameters of CP. The lower part of Table 2 shows that RAPS achieves a consistent performance boost when using CP pre-trained model. It is worth to mention that RAPS (CP) has 0.97 and 1.24 points of improvement compared with HCRP (CP), under the 5-way 1-shot and 10-way 1-shot scenarios, respectively, which demonstrates that RAPS is more suitable for few-shot scenarios.

RAPS also achieves a competitive performance compared with HCRP on FewRel 2.0. As shown in Table 3, RAPS is also transferable to the domain adaptation setting since RAPS uniformly outperforms HCRP with CP as the encoder. The possible reason why RAPS is worse than HCRP when using BERT as encoder is that the FewRel 2.0 with domain adaptation setting only provides the name of relations without a specific description, which has a relatively harmful impact on the adaptive fusion procedure (4) to generate a strong relation representation for the relation prototypes. Nevertheless, HCRP seems to be more robust to the lack of relation description due to the complex hybrid prototype learning module, which can learn strong contextualized embeddings even without the relation information.

4.4 Effects of Different Adaptive Prototype Fusion Methods

In this subsection, we further explore four different types of adaptive prototype fusion methods

- Unconstrained Adaptive Scalar (UAS): $\mathbf{p}_i^w = w_1\mathbf{p}_i + w_2\mathbf{r}_i, w_1, w_2 \in \mathbb{R}$.
- Constrained Adaptive Scalar (CAS): $\mathbf{p}_i^w = w_1\mathbf{p}_i + w_2\mathbf{r}_i, w_1, w_2 \in \mathbb{R}$, subject to $w_1 + w_2 = 1$.
- Unconstrained Adaptive Matrix (UAM): $\mathbf{p}_i^W = \mathbf{W}_1\mathbf{p}_i + \mathbf{W}_2\mathbf{r}_i, \mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{2d\times 2d}$.
- Constrained Adaptive Matrix (CAM): $\mathbf{p}_i^W = \mathbf{W}_1\mathbf{p}_i + \mathbf{W}_2\mathbf{r}_i, \mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{2d\times 2d}$, subject to $\mathbf{W}_1 + \mathbf{W}_2 = \mathbf{I}_{2d}$.

For UAS and CAS, w_1 and w_2 are two learnable scalars. For UAM and CAM, $\mathbf{W}_1$ and $\mathbf{W}_2$ are two learnable matrices, and $\mathbf{I}_{2d}$ is the identity matrix of order $2d$. Note that UAM and CAM are equivalent to adding a fully-connected layer without bias term on $\mathbf{p}_i$ and $\mathbf{r}_i$, respectively. We evaluate these four types of adaptive prototype fusion methods on the FewRel 1.0 validation set. The corresponding results are shown in Table 4. We can see that UAS performs the best in all scenarios. Compared with CAS, UAS has more degrees of freedom to learn the trade-off between the relation prototype and relation information, thus leading to more accurate prototype and better performance than CAS. It can be seen that the matrix-based fusion methods (UAM and CAM) are consistently inferior to scalar-based fusion methods (UAS and CAS), which may attribute to the introduction of many useless parameters ($O(d^2)$ for UAM and CAM vs. $O(1)$ for UAS and CAS).

[2] https://codalab.lisn.upsaclay.fr

Model	Mean	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
UAS	90.86	92.26	94.08	87.23	89.87
CAS	89.79	91.91	93.09	86.55	87.61
UAM	75.02	78.56	85.84	64.58	71.11
CAM	73.58	74.64	84.69	61.37	73.63

Table 4: Results of different prototype fusion methods under different scenarios on the FewRel 1.0 validation set (%).

4.5 Ablation Studies

In this subsection, we conduct an ablation study on 5-way 1-shot and 10-way 1-shot scenarios with BERT based on the FewRel 1.0 validation set, to demonstrate the effectiveness of the proposed query-information guided attention module and adaptive prototype fusion mechanism (abbreviated as QIA and APF for further reference, respectively). We consider three ablation experiments including w/o QIA, w/o APF, and w/o QIA and APF.

- w/o QIA: calculate the prototype using (1) instead of (2) + (3).
- w/o APF: calculate $\mathbf{p}_i^w$ using direct addition $\mathbf{p}_i^w = \mathbf{p}_i + \mathbf{r}_i$ instead of (4).
- w/o QIA and APF: calculate the prototype using (1) instead of (2) + (3), and calculate $\mathbf{p}_i^w$ using direct addition $\mathbf{p}_i^w = \mathbf{p}_i + \mathbf{r}_i$ instead of (4).

From the results in Table 5, We can obtain several observations. First, the model performance drops more or less without any of QIA or APF, which demonstrates the effectiveness of RAPS. Second, QIA seems to be more important under 10-way 1-shot scenario than 5-way 1-shot, with accuracy losses of 0.15% and 0.76%, respectively. Third, APF seems to be more important under 5-way 1-shot scenario than 10-way 1-shot, with accuracy losses of 1.35% and 1.13%, respectively. Based on the above discussion, it may be a future research direction on how to choose appropriate prototype computation method under different N-way K-shot settings.

Method	5-way 1-shot	10-way 1-shot
QIA+APF	92.26	87.23
w/o QIA	92.11	86.47
w/o APF	91.91	86.10
w/o QIA and APF	91.33	86.02

Table 5: Results of ablation study on FewRel 1.0 validation set (%). w/o is the abbreviation of without.

4.6 Visualization

4.6.1 Visualization of Queries

To further demonstrate the excellent discriminative capability of RAPS, we present some t-SNE (van der Maaten and Hinton, 2008) visualization results with BERT on the 5-way 1-shot meta task based on the validation set of FewRel 1.0 and FewRel 2.0, respectively. Figure 2 and Figure 3 show the t-SNE visualization of query instances, where different colors represent different relation classes. The left subfigures show the statements of query instances encoded by Proto-BERT (Snell et al., 2017), and the right subfigures show the statements of query instances encoded by RAPS.

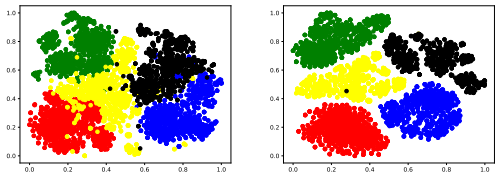

Figure 2: Visualization of queries from FewRel 1.0 validation set. Left: Instances encoded by Proto-BERT; Right: Instances encoded by RAPS.

It can be seen from the left subfigure of Figure 2 that, although the instances can also be divided into different classes, the intra-class instances are not close enough, and there are multiple error points (i.e., yellow points in red cluster). After introducing the relation information and training with the proposed query-information guided attention and adaptive prototype fusion mechanism, we can see from the right subfigure that the representations of the same class are closer, and the inter-class boundaries are more distinct. Meanwhile, there seems to be only one error point. As for Figure 3, the intra-class distances are wider than the counterpart in Figure 2, because the model is trained on Wikipedia data of universal domain but inferences on the PubMed data of biomedical domain, which may lead to the difficulty in learning good intra-class similarity. However, the inter-class boundaries in the right subfigure are still distinct enough to distinguish the inter-class dissimilarity. The observation shows that RAPS is indeed beneficial to the improvement of the model.

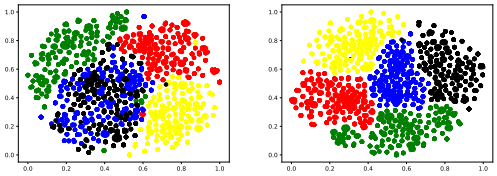

Figure 3: Visualization of queries from FewRel 2.0 validation set. Left: Instances encoded by Proto-BERT; Right: Instances encoded by RAPS.

4.6.2 Visualization of Easy-to-confuse Relations

As Figure 4 depicts, we visualize the learned embedding spaces with t-SNE to intuitively characterize the resulting representations for similar relations. Specifically, we pick three similar relations including "mother" (P25), "spouse" (P26) and "child" (P40) from the FewRel 1.0 validation set. We can see that embeddings of three easy-to-confuse relations are clearly separated, which makes classification easier. Even for hard tasks, our model still arrives at discriminative representations.

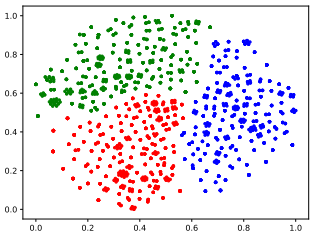

Figure 4: t-SNE plot of instance embeddings. Three easy-to-confuse relations (“mother”, "spouse" and “child”) are adopted.

5 Conclusion

In this paper, we propose a novel FSRE pipeline using the proposed query-information guided attention module and adaptive prototype fusion mechanism, called RAPS. It has two advantages: 1) RAPS exploits rich interactive information between support instances and query instances to obtain more accurate initial relation prototypes. 2) RAPS dynamically makes a trade-off between the derived relation prototype and relation information by the adaptive prototype fusion mechanism to obtain more compatible integrated relation prototype. Experiment results on FewRel 1.0 and FewRel 2.0 datasets show that RAPS achieves a significant improvement against the modern state-of-the art methods. One possible direction of future work is to generalize the adaptive prototype fusion mechanism to other text classification tasks such as intent classification and sentiment analysis.

Limitations

There are several limitations of this work. First, RAPS only works under the N-way K-shot setup, because it requires a support set to calculate the relation class prototype. Second, its effectiveness is only examined on the task of few-shot relation extraction, while the generalization to other text classification tasks, such as intent classification and news classification, is not yet explored in this paper. Third, the model has not been extended to perform non-of-the-above (NOTA) detection (Gao et al., 2019b), where a query instance may not belong to any class in the support set.

References

Nguyen Bach and Sameer Badaskar. 2007. A review of relation extraction. Literature Review for Language and Statistics II, 2:1–15.

Trapit Bansal, Karthick Prasad Gunasekaran, Tong Wang, Tsendsuren Munkhdalai, and Andrew McCallum. 2021. Diverse distributions of self-supervised tasks for meta-learning in NLP. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5812–5824, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xin Cong, Bowen Yu, Tingwen Liu, Shiyao Cui, Hengzhu Tang, and Bin Wang. 2021. Inductive unsupervised domain adaptation for few-shot classification via clustering. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 624–639. Springer.

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.

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive graph for multi-hop reading comprehension at scale. In Proceedings of the 57th Annual Meeting of the Association for

Computational Linguistics, pages 2694–2703, Florence, Italy. Association for Computational Linguistics.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dheeru Dua, Sameer Singh, and Matt Gardner. 2020. Benefits of intermediate annotations in reading comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5627–5634, Online. Association for Computational Linguistics.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning, pages 1126–1135. PMLR.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019a. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In The Thirty-Third AAAI Conference on Artificial Intelligence, 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 6250–6255, Hong Kong, China. Association for Computational Linguistics.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Jinpeng Hu, Jianling Li, Zhihong Chen, Yaling Shen, Yan Song, Xiang Wan, and Tsung-Hui Chang. 2021. Word graph guided summarization for radiology findings. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4980–4990, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.

Zhifei Li, Hai Liu, Zhaoli Zhang, Tingting Liu, and Neal N. Xiong. 2022. Learning knowledge graph embedding with heterogeneous relation attention networks. IEEE Transactions on Neural Networks and Learning Systems, 33(8):3961–3973.

Songtao Liu, Huang Di, and Wang Yunhong. 2019. Learning spatial fusion for single-shot object detection. arxiv preprint arXiv:1911.09516.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In International Conference on Learning Representations.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

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, Online. Association for Computational Linguistics.

Meng Qu, Tianyu Gao, Louis-Pascal A. C. Xhonneux, and Jian Tang. 2020. Few-shot Relation Extraction via Bayesian Meta-learning on Relation Graphs. In Proceedings of the 37th International Conference on Machine Learning, pages 7867–7876. PMLR.

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

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, volume 30, pages 4077–4087. Curran Associates, Inc.

Baldini Livio 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.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research, 9(11).

Mengru Wang, Jianming Zheng, Fei Cai, Taihua Shao, and Honghui Chen. 2022. DRK: Discriminative rule-based knowledge for relieving prediction confusions in few-shot relation extraction. In Proceedings of the 29th International Conference on Computational Linguistics, pages 2129–2140, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yingyao Wang, Junwei Bao, Guangyi Liu, Youzheng Wu, Xiaodong He, Bowen Zhou, and Tiejun Zhao. 2020. Learning to decouple relations: Few-shot relation classification with entity-guided attention and confusion-aware training. In Proceedings of the 28th International Conference on Computational Linguistics, pages 5799–5809, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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 amp; Knowledge Management, CIKM '20, page 2273–2276, New York, NY, USA. Association for Computing Machinery.

Shan Yang, Yongfei Zhang, Guanglin Niu, Qinghua Zhao, and Shiliang Pu. 2021. Entity concept-enhanced few-shot relation extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 987–991, Online. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Linhai Zhang, Chao Lin, Deyu Zhou, Yulan He, and Meng Zhang. 2021. A Bayesian end-to-end model with estimated uncertainties for simple question answering over knowledge bases. Computer Speech & Language, 66:101167.

Yu Zhao, Anxiang Zhang, Ruobing Xie, Kang Liu, and Xiaojie Wang. 2020. Connecting embeddings for knowledge graph entity typing. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6419–6428, Online. Association for Computational Linguistics.