Enhancing the Prototype Network with Local-to-Global Optimization for Few-Shot Relation Extraction

Hui Sun[1], **Rongxin Chen**[1†]
[1]College of Computer Engineering, Jimei University, China
`sylvan@jmu.edu.cn, ch2002star@163.com`

Abstract

Few-Shot Relation Extraction (FSRE) aims to achieve high classification performance by training relation classification models with a small amount of labeled data. Prototypical networks serve as a straightforward and efficient method for optimizing model performance by combining similarity evaluation and contrastive learning. However, directly integrating these methods can introduce unpredictable noise, such as **information redundancy**, which hinders classification performance and negatively affects embedding space learning. The technique presented in this paper applies **Lo**cal-**To**-**G**lobal optimization to enhance prototypical networks in few-shot relation extraction. Specifically, this paper develops a local optimization strategy that indirectly optimizes the prototypes by optimizing the other information contained within the prototypes. It considers relation prototypes as global anchors and incorporates the techniques introduced in this paper, such as information alignment, local contrastive learning, and a local adaptive focal loss function, to address the issues of information redundancy. This approach enables the model to learn a unified and effective embedding space. We conduct extensive experiments on the FewRel 1.0 and FewRel 2.0 datasets to validate the effectiveness of the proposed model[1].

1 Introduction

Relation Extraction (RE) is a core task in Natural Language Processing (NLP) that aims to automatically identify and extract semantic relationships between entities from unstructured text. The objective of RE is to convert natural language knowledge into an organized format that computers can comprehend and process. RE is widely used in many different NLP tasks, including knowledge graph construction (Zhang et al., 2019), machine reading comprehension (Abdou et al., 2019), question answering (Saxena et al., 2020), and dialogue systems (Ma et al., 2019). However, the annotation of large-scale datasets often demands significant human and material resources, and the lack of sufficient data in smaller datasets limits model performance. In recent years, Few-Shot Relation Extraction (FSRE) has emerged as an effective solution to the challenges of annotation difficulty and data scarcity. The objective of FSRE is to train models using a small amount of labeled data and enable them to quickly adapt to new types of relationships, even when data is extremely limited.

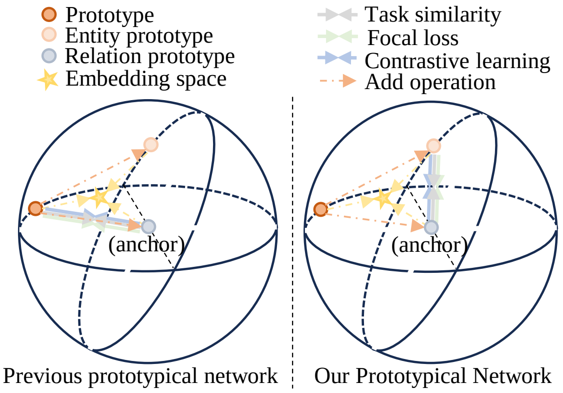


Figure 1: Illustration of the difference between previous Prototypical Network model and our Prototypical Network. (Best Viewed in Color).

Prototype networks (Snell et al., 2017) stand out as a straightforward but efficient method among the various FSRE algorithms. Prototypical networks have the primary goal of classifying query instances through the use of prototypes. Prototypes for each relation class are learned using support instances from the support set. To be more precise, the model takes a new query sample, calculates its distance from the learned prototypes, and then assigns the sample to the relation class of the closest prototype. Recent work has included relation descriptions into the model (Yang et al., 2020; Han

[†]Corresponding author.

[1]`https://github.com/sunxingzheowo/LoToG`

Findings of the Association for Computational Linguistics: NAACL 2025, pages 2668–2677
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

et al., 2021; Liu et al., 2022) to improve the quality of the learned embeddings and further improve the performance of prototypical networks.

Recent research has introduced various contrastive learning strategies to mitigate prediction confusion between similar classes. For instance, AdapAug (Li et al., 2024) proposed an adaptive class prototype network that employs both instance-level and representation-level augmentation. This method uses a perturbation attention mechanism in conjunction with a gradual addition of new instances to accomplish contrastive training on relational prototypes. Alternately, by aligning multiple representations and extracting discriminative information complementary to each representation, MultiRep (Borchert et al., 2024) is an information extraction technique that maximizes model performance. It does this by utilizing contrastive learning and multi-sentence representations. HCRP (Han et al., 2021) introduced a hybrid contrastive relation prototype method that uses relational information as an anchor. This method treats samples from different classes as negative examples and samples from the same class within a prototype as positive examples. Through contrastive learning, the method brings instances of the same class closer together and pushes dissimilar instances apart. However, despite the promising results, this method overlooks the relationship between prototype information and relational information. Specifically, when relational information is embedded within the prototype and used as an anchor, calculating similarity or performing contrastive learning with respect to this anchor may lead to other important information within the prototype being weakened or overlooked during training. This can negatively impact the learned embedding space from the prototype. We refer to this phenomenon as **information redundancy**, which occurs when objects share overlapping components, leading to the generation of duplicate or redundant information. From a computational perspective, we argue that information redundancy may cause the model to overemphasize the shared components while neglecting the unique or independent aspects, thereby diminishing the effectiveness of the representations.

In order to address the issue of information redundancy within prototypes, this paper suggests a way to improve prototypical networks by switching from local to global optimization. In particular, we align relational information with entity information to align the head and tail entities with the same relational information in the spatial representation before generating both relation and entity prototypes. Building on SimpleFSRE's perspective (Liu et al., 2022), we combine relation prototypes and entity prototypes into a single prototype representation. We then separately analyze the relation and entity prototypes within the overall prototype structure. In the entity prototype, instances from the same class serve as positive samples, while instances from different classes serve as negative samples. Global anchoring is thought to apply to the relation prototype. Through contrastive learning, the entity prototype learns its spatial representation from the relation prototype. The model can learn a more efficient embedding space without undervaluing the entity prototype by assembling the prototype from both relation and entity prototypes.

Furthermore, we introduce a local adaptive focal loss function. Since the prototype contains the relation prototype, we do not directly compute the similarity between the overall prototype and the relation prototype. Rather than employing a balancing hyperparameter in the focal loss function, we compute the similarity between the entity prototype and the relation prototype. By taking this approach, the prototype's embedding space is indirectly optimized, which effectively reduces redundant information and improves the discriminative power of the model.

The method proposed in this paper can be intuitively understood as illustrated in Figure 1. By employing local optimization, it maximizes the utilization of each component of the prototype, alleviating the negative effects of information redundancy and ultimately enhancing the performance of the prototypical network. In summary, our contributions are:

1) We have proposed a novel method that employs local optimization strategies to enhance the prototype representation through indirect optimization.

2) We have introduced an information alignment strategy that aligns entity and relational information, enabling both head and tail entities to learn consistent relational information via contrastive learning.

3) We have developed a local contrastive learning approach that mitigates the effects of re-

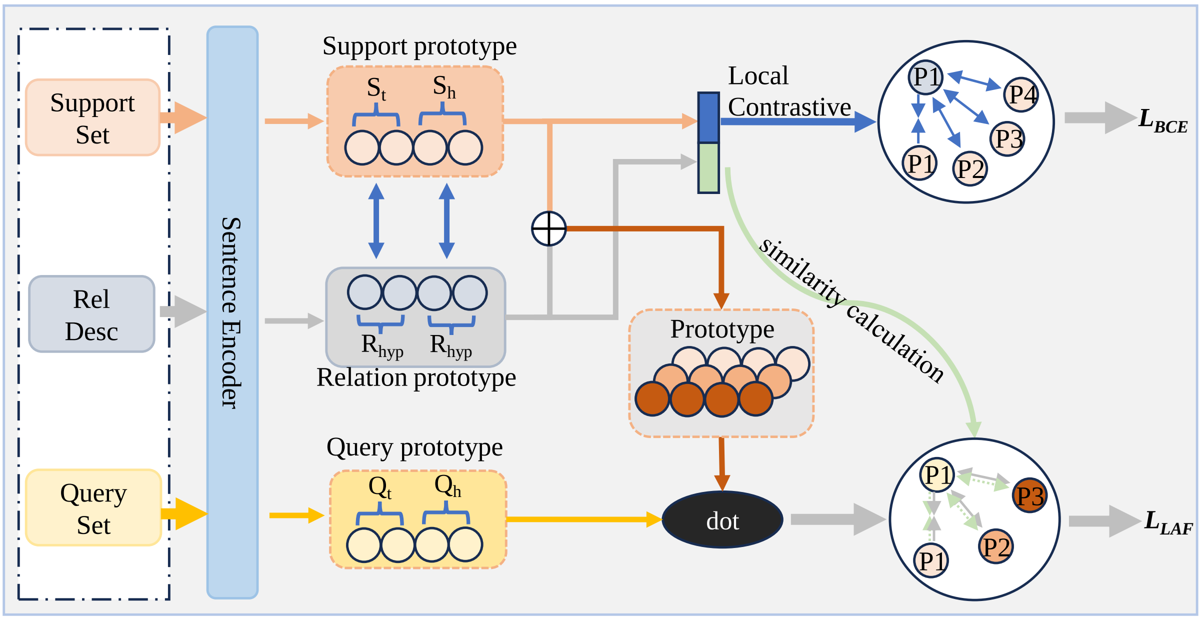


Figure 2: The overall framework of LotoG (Best viewed in color). The relation stream, support set stream, and query set stream are represented in gray, orange, and yellow, respectively.⊕ denotes the addition operation.

dundant information, facilitating the indirect optimization of the prototype.

4) We have introduced a local adaptive focal loss function that refines the prototype representation by leveraging local similarity for improved optimization.

2 Methodology

In this section, we introduce a Local-to-Global Optimization (LoToG) framework designed to enhance prototypical networks for few-shot learning. This framework leverages relational information as a global anchor point, allowing for the optimization of individual prototype components, which in turn facilitates the indirect optimization of the overall prototype. By addressing the noise introduced by the information redundancy, the framework enables the model to achieve a more effective embedding space. As illustrated in Figure 2, LoToG comprises five key components:

(1) Encoding Text Information: The textual data is encoded into an embedded representation. (2) Information Alignment: The head and tail entities are aligned with the same relational information. Through contrastive learning, entities learn shared relational information. (3) Local Contrastive Learning: The remaining components of the prototype, specifically the entity prototypes, engage in contrastive learning with the relation prototype (anchor), which helps reduce noise arising from information redundancy. (4) Local Adaptive Focal Loss Function: The similarity between the entity prototypes and the anchor serves as a balancing factor. When combined with the focal loss function, this enhances the optimization of the prototype representation, thereby further improving the prototypical network through local optimization. (5) Prediction: The similarity between the overall prototype and the query prototype is computed to ascertain the corresponding class label.

2.1 Task Definition

We follow a typical few-shot task setting, known as the N-way-K-shot setting, which includes a support set $\boldsymbol{S}$ and a query set $\boldsymbol{Q}$. The support set $\boldsymbol{S}$ consists of N different classes, with each class having K labeled instances. The query set $\boldsymbol{Q}$ contains the same N classes as $\boldsymbol{S}$, and the task is evaluated on the query set $\boldsymbol{Q}$, attempting to predict the class to which each relation in $\boldsymbol{Q}$ belongs. We utilized two datasets, where the training set contains abundant base classes, each with a large number of labeled examples, and the validation set contains various novel classes, each with a certain number of labeled examples. Note that base classes and novel classes are disjoint. Few-shot learning involves acquiring knowledge from base classes and utilizing this knowledge to identify novel classes. Specifically, in each training iteration, we randomly select N classes from the base classes, with each class having K instances, forming the support set $\boldsymbol{S} = \{\boldsymbol{s}_k^i \mid i = 1, 2, \ldots, N;\ k = 1, 2, \ldots, K\}$. At the same time, we extract M instances from the remaining data of N classes to construct the query set $\boldsymbol{Q} = \{\boldsymbol{q}_m^i \mid m = 1, 2, \ldots, M\}$. For the FSRE task, each instance consists of a tuple (x, e, y), where x represents a natural language sentence, $e = (e_h, e_t)$ represents a pair of head entity and tail entity, and y is the relationship label.

2.2 Sentence Encoder

We utilize BERT (Devlin et al., 2019) as our encoder to derive contextual embeddings for both the support set $\boldsymbol{S}$ and the query set $\boldsymbol{Q}$. Specifically, for the global representation of each instance, we first use a special token to mark the positions of the head and tail entities within the instance. The BERT encoder is then used to obtain the contextual representation of the starting position, which generates the head entity representation $\boldsymbol{s}_{k,h}^i$ and the tail entity representation $\boldsymbol{s}_{k,t}^i$ for the support set, while the query set generates the corresponding head entity representation $\boldsymbol{q}_{m,h}^i$ and tail entity representation $\boldsymbol{q}_{m,t}^i$. Additionally, the relation description is encoded to produce a relation representation with the same dimensionality as the instance representation. Specifically, the relation sentence is input as "[CLS] Relation Name [SEP] Relation Description," and the global representation $\boldsymbol{r}_g^i$ is obtained from the [CLS] token's representation. For the local information representation of each relation, we first tokenize the instance using the WordPiece tokenizer, splitting words into subwords. We then directly extract all hidden states from the final layer of the BERT model to obtain the representations of all subwords. The maximum length of an instance is denoted as L, indicating that each relation contains L units of local information, generating the corresponding local representation $\boldsymbol{r}_l^i$. To facilitate the representation of relationships among these entities, we will refer to the following terms in subsequent discussions to denote their corresponding features: $\{\boldsymbol{r}_l^i, \boldsymbol{r}_g^i, \boldsymbol{s}_{k,h}^i, \boldsymbol{s}_{k,t}^i, \boldsymbol{q}_{m,h}^i, \boldsymbol{q}_{m,t}^i\} \in \mathbb{R}^d$, where $\mathbb{R}^d$ represents the feature dimensions obtained from the sentence encoder.

2.3 Representation space alignment

To effectively map entity information and relational information into a unified representation space, we propose an information alignment strategy. This strategy aligns the head entity and tail entity information produced by the BERT encoder with the corresponding relational information it generates. Such alignment enhances subsequent contrastive learning, enabling both head and tail entities to share a common representation space.

As detailed in Section 2.2, we utilize $\boldsymbol{r}_l^i$ and $\boldsymbol{r}_g^i$ to represent the relational information. Given that we treat relational information as a global anchor, we first extract the global relational information $\boldsymbol{r}_g^i$ and the local relational information $\boldsymbol{r}_l^i$, subsequently combining them to create a hybrid relation prototype $\boldsymbol{r}_{hyp}^i$. Finally, we concatenate the two $\boldsymbol{r}_{hyp}^i$ prototypes to derive the final relation prototype $\boldsymbol{r}_p^i$ (serving as the global anchor). This can be formulated as follows:

$$\boldsymbol{r}_{hyp}^i = \boldsymbol{r}_l^i + \boldsymbol{r}_g^i \in \mathbb{R}^{2d}, \tag{1}$$

$$\boldsymbol{r}_p^i = \begin{bmatrix} \boldsymbol{r}_{hyp}^i \\ \boldsymbol{r}_{hyp}^i \end{bmatrix} \in \mathbb{R}^{2d}, \tag{2}$$

Based on the representations outlined in Section 2.2, and following the information alignment strategy, we concatenate $\boldsymbol{s}_{k,h}^i$ and $\boldsymbol{s}_{k,t}^i$ together and then compute the average across K support samples of the same class to form the entity prototype $\boldsymbol{s}_p^i$ for the support set. For the query set, we simply concatenate the head and tail entities to form the query entity prototype $\boldsymbol{q}_{m,p}^i$. The main objective is to align these head and tail entities with the hybrid relation prototype. This can be formally expressed as follows:

$$\boldsymbol{s}_p^i = -\frac{1}{K} \sum_{k}^{K} \begin{bmatrix} \boldsymbol{s}_{k,h}^i \\ \boldsymbol{s}_{k,t}^i \end{bmatrix} \in \mathbb{R}^{2d}, \tag{3}$$

$$\boldsymbol{q}_{m,p}^i = \begin{bmatrix} \boldsymbol{q}_{m,h}^i \\ \boldsymbol{q}_{m,t}^i \end{bmatrix} \in \mathbb{R}^{2d}. \tag{4}$$

2.4 Local contrastive learning

As described in Section 2.3, we derive the relation prototype $\boldsymbol{r}_p^i$ and the entity prototypes $\boldsymbol{s}_p^i$. In this subsection, we focus on the proposed local contrastive learning approach. Specifically, we begin by performing an addition operation between the relation prototype and the instance prototypes to obtain the prototype representation $\boldsymbol{P} = \{\boldsymbol{p}^i\}$, as illustrated below:

$$\boldsymbol{p}^i = \boldsymbol{r}_p^i + \boldsymbol{s}_p^i \in \mathbb{R}^{2d}, \tag{5}$$

N-way-K-shot	5-way-1-shot		5-way-5-shot		10-way-1-shot		10-way-5-shot	
Models (Encoder: BERT)	val	test	val	test	val	test	val	test
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
BERT-PAIR (Gao et al., 2019)	85.66	88.32	89.48	93.22	76.84	80.63	81.76	87.02
MTB (Baldini Soares et al., 2019)	–	91.10	–	95.40	–	84.30	–	91.80
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
HCRP (Han et al., 2021)	90.90	93.76	93.22	95.66	85.11	89.95	87.79	92.10
CTEG (Wang et al., 2020)	84.72	88.11	92.52	95.25	76.01	81.29	84.89	91.33
SimpleFSRE (Liu et al., 2022)	91.29	94.42	94.05	96.37	86.09	90.73	89.68	**93.47**
DAPL (Yu et al., 2022)	–	85.94	–	94.28	–	77.59	–	89.26
LPD (Zhang and Lu, 2022)	88.84	93.79	90.65	95.07	79.61	89.39	82.15	91.08
FAEA (Dou et al., 2022)	90.81	95.10	94.24	96.48	84.22	90.12	88.74	92.72
AdapAug (Li et al., 2024)	92.27	94.35	93.95	**96.96**	86.01	90.69	89.67	93.46
MultiRep (Borchert et al., 2024),	**92.73**	94.18	93.79	96.29	86.12	91.07	88.80	91.98
Ours(LoToG)	92.38	**95.28**	**94.26**	96.71	**86.23**	**91.48**	**91.11**	93.14

Table 1: FewRel 1.0 validation/testing set few-shot classification accuracy (%). The best results are presented in bold, while the second-best results are underlined..

Previous contrastive learning methods (Han et al., 2021) directly compared the prototype with the anchor point. However, during the contrastive learning process, we do not split $\boldsymbol{p}^i$ into positive and negative prototypes for triplet contrastive learning with the anchor point. Since $\boldsymbol{p}^i$ already contains $\boldsymbol{r}_p^i$, we divide $\boldsymbol{s}_p^i$ into positive and negative prototypes. Samples belonging to the same class as the anchor are designated as positive prototypes, while those from different classes are designated as negative prototypes. The model collects the positive prototypes $\boldsymbol{s}_p^i$ and the negative prototypes $\{\boldsymbol{s}_p^n; n = 1, \ldots, N, n \neq i\}$. The goal is to distinguish between positive and negative prototypes. Given the true label $\{\mathcal{T}_{ij}; j = 1, \ldots, N\}$, we employ the dot product $\mathcal{G}(\cdot)$ to compute the logits between the anchor and the selected prototypes, denoted as $\mathcal{P}_{ij}$, as follows.

$$\mathcal{T}_{ij} = \begin{cases} 1, & \mathcal{G}(\boldsymbol{r}_p^i \cdot \boldsymbol{s}_p^i)\ j = i; \\ 0, & \mathcal{G}(\boldsymbol{r}_p^i \cdot \boldsymbol{s}_p^n)\ j = n, \end{cases} \tag{6}$$

$$\mathcal{P}_{ij} = \mathcal{G}(\boldsymbol{r}_p^i \cdot \boldsymbol{s}_p^j), \tag{7}$$

After categorizing positive and negative samples from $\boldsymbol{S}_p = \{\boldsymbol{s}_p^j\}$, we reformulate the problem as a binary classification task. We then perform contrastive learning with the anchor point to ensure that entity representations belonging to the same class are positioned closer to the anchor point, while those from different classes are placed farther away. The prototype derived from the addition operation preserves both relational and entity information, effectively mitigating the influence of redundant information. For the binary classification task, we employ the Binary Cross-Entropy Loss with Logits, denoted as $\mathcal{L}_{BCE}$. This method integrates the sigmoid activation function with binary cross-entropy loss in a numerically stable fashion. The loss function is defined as follows:

$$\begin{aligned} \mathcal{L}_{BCE} = &- [y \log(\sigma(a)) \\ &+ (1-y)\log(1-\sigma(a))], \end{aligned} \tag{8}$$

where a denotes the output logits of the model, y represents the true label (0 or 1), and $\sigma(a)$ is the probability derived by applying the sigmoid function to the logits a.

Therefore, the loss function is computed as follows:

$$\begin{aligned} \mathcal{L}_{BCE} = &- \frac{1}{N^2} \sum_{i}^{N} \sum_{j}^{N} \Big(\mathcal{T}_{ij} \log(\sigma(\mathcal{P}_{ij}) \\ &+ (1-\mathcal{T}_{ij}) \log(1-\sigma(\mathcal{P}_{ij}))\Big). \end{aligned} \tag{9}$$

2.5 Local adaptive focal loss

The focal loss introduced by (Lin et al., 2017) addresses the imbalance between hard and easy examples in classification tasks. Subsequently, (Han et al., 2021) enhanced this approach specifically for few-shot relation extraction tasks. The formula for the adaptive focal loss function is presented as follows:

$$\mathcal{L}_{AF} = -\alpha(1-z_y)^{\gamma} \log z_y. \tag{10}$$

The parameter γ is a difficulty-adjusting coefficient that controls the model's attention to misclassified samples. The variable y represents the class label, while z_y denotes the estimated probability for class y. The similarity factor α is determined by the degree of dispersion between classes in a batch; lower dispersion corresponds to higher class similarity, indicating that the batch should receive more focus.

Redundant information can pose challenges due to the inclusion of anchor information within the prototype; therefore, our objective is to eliminate this noise. Inspired by the adaptive focal loss framework presented in (Han et al., 2021) and addressing the issue of information redundancy, this paper introduces a local adaptive focal loss function. This adaptation enables the model to focus on more challenging instances while mitigating the impact of noise introduced by redundant information. This method utilizes the similarity between $\boldsymbol{s}_p^i$ and the anchor as the local similarity factor α_l, rather than employing the similarity between $\boldsymbol{p}^i$ and the anchor as the similarity factor α. By optimizing local similarity, we indirectly enhance the embedding space of $\boldsymbol{p}^i$. This method optimizes the entire model through local optimization without adversely affecting the overall performance of the model.

Specifically, we first perform a concatenation operation between $\boldsymbol{s}_p^i$ and $\boldsymbol{r}_p^i$ to obtain the feature vector $\boldsymbol{F}$. Next, we normalize the $\boldsymbol{F}$ vector to ensure consistent feature scales, which enhances the stability and performance of model training. Then, we compute the task similarity matrix using matrix multiplication:

$$\boldsymbol{F} = | \begin{bmatrix} \boldsymbol{s}_p^i \\ \boldsymbol{r}_p^i \end{bmatrix} |, \tag{11}$$

where $|\cdot|$ denotes the Euclidean norm. The task similarity scalar is calculated as follows:

$$\alpha_l = \text{softmax}(||\boldsymbol{F} \cdot \boldsymbol{F}^\top||), \tag{12}$$

where $||\cdot||$ denotes the Frobenius norm. The local adaptive loss function is defined as follows:

$$\mathcal{L}_{LAF} = -\alpha_l (1 - z_y)^\gamma \log z_y, \tag{13}$$

Therefore, the loss function is computed as follows:

$$(z_y)_m^{i,j} = \text{softmax}(\mathcal{G}(\boldsymbol{q}_{m,p}^i \cdot \boldsymbol{p}^j)), \tag{14}$$

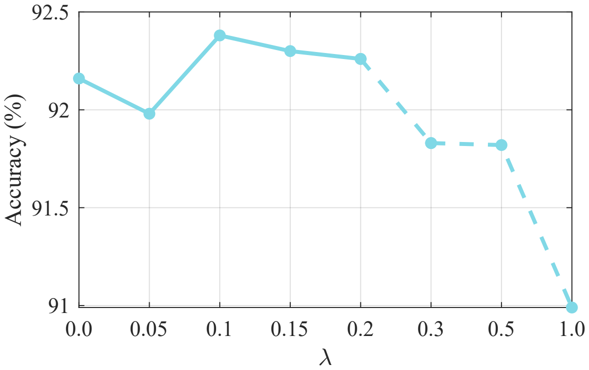


Figure 3: Accuracy for different λ values on the FewRel 1.0 validation set for the 5-way-1-shot task.

$$\mathcal{L}_{LAF} = -\frac{1}{MN^2} \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{j=1}^{N} \frac{\alpha_l^i (1 - (z_y)_m^{i,j})^\gamma \log((z_y)_m^{i,j})}{\left(\sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{j=1}^{N} \alpha_l^i (1 - (z_y)_m^{i,j})^\gamma\right)}. \tag{15}$$

α_l^i represents the local dynamic adjustment factor that dynamically modifies the loss based on the difficulty associated with class i. The term $(z_y)_m^{i,j}$ denotes the estimated probability between the m-th query sample of class i and the j-th class prototype.

The final objective function of our model is defined as follows:

$$\mathcal{L} = \mathcal{L}_{LAF} + \lambda \times \mathcal{L}_{BCE} \tag{16}$$

where λ is a hyperparameter to balance the two terms.

3 Experiments

3.1 Datasets

Our model is evaluated on two commonly-used datasets:

1) FewRel 1.0 (Han et al., 2018). This is a large-scale, manually annotated FSRE dataset constructed from Wikipedia articles, containing 100 different relations. Each relation includes 700 instances. The dataset is split into training, validation, and test sets with 6, 4, 16, and 20 relations, respectively.

2) FewRel 2.0 (Gao et al., 2019). To evaluate the generalization capability of our model, we also conducted experiments on FewRel 2.0. The training set for FewRel 2.0 is the same as that of FewRel 1.0. Notably, the test set of FewRel 2.0 is constructed from the

Models	5-way	10-way
(Encoder: BERT)	5-shot	5-shot
Proto-BERT (Snell et al., 2017)	51.50	36.39
BERT-PAIR (Gao et al., 2019)	78.57	66.85
HCRP (Han et al., 2021)	83.03	72.94
AdapAug (Li et al., 2024)	84.37	73.92
Ours(LoToG)	**84.38**	**75.69**

Table 2: We chose the more challenging FewRel 2.0 dataset and configured the 10-way-5-shot and 5-way-5-shot settings to demonstrate the effectiveness of our method.

biomedical domain, which does not overlap with the relations in the training set. It contains 25 relations, each with 100 instances. FewRel 2.0 is trained on the Wikipedia domain but tested on a different biomedical domain. Only the names of the relation labels are provided, without descriptions, making the task more challenging.

3.2 Analysis of the Hyperparameter λ

In equation 16, the hyperparameter λ is used to balance the two loss terms. To determine the optimal value of λ, we conducted experiments with various λ settings on the FewRel 1.0 validation set. As shown in Figure 3, the model achieved the best performance when $\lambda = 0.1$. Therefore, we employed $\lambda = 0.1$ in all subsequent FewRel 1.0 experiments. Given the specific characteristics of the FewRel 2.0 dataset, we set $\lambda = 1$ for all subsequent experiments on FewRel 2.0.

3.3 Setup

We adopt the BERT-base-uncased model from Huggingface's Transformer[2](Wolf et al., 2020) library as the encoder, which contains approximately 110 million parameters. The implementation is based on PyTorch (Paszke et al., 2019). The AdamW optimizer (Loshchilov and Hutter, 2019) is employed to minimize the loss. Following the training and evaluation procedures outlined by (Gao et al., 2019), our model was trained for 30,000 iterations on the FewRel training set, with 1,000 iterations for validation and 10,000 iterations for testing. The batch size was set to 4, and the learning rate was configured at $2e-5$. The total training time for the four types of N-way-K-shot tasks on an NVIDIA RTX 4090D GPU is approximately 16 hours, while testing takes about 3 hours.

Models	5-way	10-way
(Encoder: BERT)	1-shot	1-shot
Ours (LoToG)	**92.38**	**86.23**
$w/o.$ Representation alignment	90.89	84.32
$w/o.$ Local contrastive	92.16	85.98
$w/o.$ $\mathcal{L}_{LAF}$	91.96	85.37
$w/.$ $\mathcal{L}_{AF}$ & proto contrastive	91.00	83.26

Table 3: Ablation results on the FewRel 1.0 validation set. "$w/o.$ x": Without module x. "$w/.$ x": With module x.

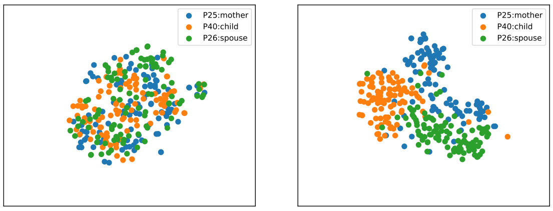
3.4 Comparison with Baselines

We compare our model with the following baseline methods: 1) **Proto-BERT** (Snell et al., 2017), A prototype network model based on BERT. 2) **MAML** (Finn et al., 2017), A typical meta-learning method. 3) **BERT-PAIR** (Gao et al., 2019), A similarity-based prediction method, where each query instance is paired with all support instances. 4) **MTB** (Baldini Soares et al., 2019), A BERT-based model further pre-trained with an additional matching-the-blank objective. 5) **REGRAB** (Qu et al., 2020), A method based on relation graphs. 6) **TD-Proto** (Yang et al., 2020), A prototype network model enhanced by entity descriptions. 7) **HCRP** (Han et al., 2021), An improved Proto-BERT with a hybrid attention module and task-adaptive focus loss. 8) **CTEG** (Wang et al., 2020), Learning from unlabelled data for clinical semantic textual similarity. 9) **SimpleFSRE** (Liu et al., 2022), A prototype network model enhanced by relation descriptions. 10) **DAPL** (Yu et al., 2022), A novel dependency-aware prototype learning method for few-shot relation classification. 11) **LPD** (Zhang and Lu, 2022), A label prompt dropout method that effectively utilizes relation descriptions. 12) **FAEA** (Dou et al., 2022), A functional word adaptive enhanced few-shot relation classification attention network. 13) **AdapAug** (Li et al., 2024), A adaptive class augmented prototype network for few-shot relation extraction 14) **MultiRep** (Borchert et al., 2024), A efficient information extraction in few-shot relation classification through comparative representation learning.

3.5 Results

Results on FewRel 1.0. The experimental results on the FewRel 1.0 validation and test sets are presented in Table 1. All models directly utilize BERT as the encoder without additional pre-training. Our model achieves strong performance under the same

[2] https://github.com/huggingface/transformers



(a) BERT Proto

(b) LoToG

Figure 4: The t-SNE visualization demonstrates the instance embeddings of 100 randomly sampled hard relation classification examples obtained from LoToG and the original BERT.

configurations, indicating that it demonstrates better generalization ability and stronger performance, enabling it to better handle the challenges of data scarcity in difficult few-shot scenarios.

Results on FewRel 2.0. As detailed in Section 3.1, FewRel 2.0 poses greater challenges as its training and test sets feature relation classes that are mutually exclusive and originate from distinct domains. From Table 2, it is evident that our model also performs well under the same configurations. These findings further underscore the efficacy of our model for cross-domain datasets and its enhanced generalization capability. The main results of this paper can be found in the CodaLab[3] competition.

3.6 Ablation Study

To further understand the effectiveness of the module we proposed, we conducted extensive ablation studies. Under the 5-way-1-shot and 10-way-1-shot settings, we performed ablation experiments on the FewRel 1.0 dataset by removing different modules of the model to evaluate their contributions to overall performance. The results of the ablation study are shown in Table 3. We can see that when we remove individual modules, the model's performance decreases to some extent. When the methods in (Han et al., 2021) are used to replace local contrastive learning and the local adaptive focal loss function, we observe a significant decline in the model's performance. This indicates the effectiveness of our proposed model.

3.7 Analysis of Local Optimization Effects

As shown in Figure 5, we use t-SNE to visualize the learned embedding space (van der Maaten and Hinton, 2008), providing an intuitive representation of the effects of local optimization. Specifically, we select three similar relations, "mother," "child," and "spouse," from the FewRel 1.0 validation set and randomly sample 100 instances for each relation. We visualize the prototypes produced by the original BERT output, the prototypes learned through relation-prototype contrastive learning, and compare them with the prototypes obtained using our proposed LoToG method. We observe that the embeddings trained with LoToG are more concentrated, making classification easier, while the embeddings from relation-prototype contrastive learning show relatively poor performance.

Additionally, we visualize the entity prototype embeddings in Figure 4. It can be seen that even with the optimization of the prototype embedding space, our entity prototype embeddings still exhibit clear distinctions.

4 Conclusions

This paper has presented a method for enhancing prototypical networks in few-shot relation extraction through local to global optimization. The proposed approach has effectively eliminated redundant information within the prototypes by employing relation prototypes as global anchors to optimize local information from multiple perspectives. This indirect optimization has mitigated the noise introduced by information redundancy and has enhanced the prototype embedding space, allowing each class prototype to map more effectively into a unified embedding space.

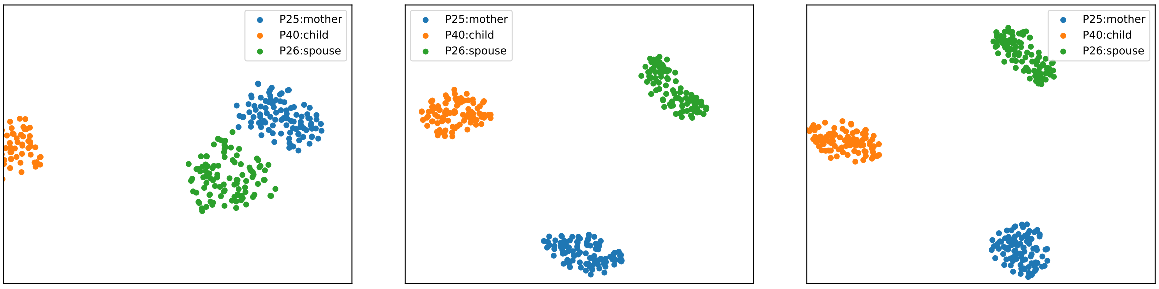
We have posited that when information redundancy is present during prototype processing, local optimization can effectively mitigate this issue without compromising the overall performance of the model. Furthermore, we have contended that these foundational principles are applicable to a diverse array of tasks that align with prototypical network modeling.

Limitations

Here are the main limitations of the proposed method:

1) We utilize relation descriptions as global anchors, which limits the applicability of our model to tasks that specifically incorporate these descriptions.

[3] https://codalab.lisn.upsaclay.fr/competitions/?q=fewrel(GoodluckGoodLuck)



(a) BERT Proto

(b) $\mathcal{L}_{AF}$ & proto contrastive

(c) LoToG

Figure 5: This figure illustrates the prototype embeddings using t-SNE under three different models, with relation descriptions included.

2) Our investigation has not encompassed the local information pertaining to entities, which is crucial for capturing the subtle distinctions between instances.

Acknowledgments

This work was supported by the Natural Science Foundation of Fujian Province of China (No.2022J01336).

References

Mostafa Abdou, Cezar Sas, Rahul Aralikatte, Isabelle Augenstein, and Anders Søgaard. 2019. X-WikiRE: A large, multilingual resource for relation extraction as machine comprehension. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 265–274, Hong Kong, China. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of ACL*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Philipp Borchert, Jochen De Weerdt, and Marie-Francine Moens. 2024. Efficient information extraction in few-shot relation classification through contrastive representation learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 638–646, Mexico City, Mexico. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chunliu Dou, Shaojuan Wu, Xiaowang Zhang, Zhiyong Feng, and Kewen Wang. 2022. Function-words adaptively enhanced attention networks for few-shot inverse relation classification. In *Proceedings of IJCAI*, pages 2937–2943. International Joint Conferences on Artificial Intelligence Organization.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of ICML*, pages 1126–1135.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of 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 EMNLP*, 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 EMNLP*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Rongzhen Li, Jiang Zhong, Wenyue Hu, Qizhu Dai, Chen Wang, Wenzhu Wang, and Xue Li. 2024. Adaptive class augmented prototype network for few-shot relation extraction. *Neural Networks*, 169:134–142.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. 2022. A simple yet effective relation information guided approach for few-shot relation extraction. In *Findings of ACL*, pages 757–763, Dublin, Ireland. Association for Computational Linguistics.

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

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, Florence, Italy. Association for Computational Linguistics.

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.

Meng Qu, Tianyu Gao, Louis-Pascal Xhonneux, and Jian Tang. 2020. Few-shot relation extraction via Bayesian meta-learning on relation graphs. In *Proceedings of ICML*, pages 7867–7876. PMLR.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4498–4507, Online. Association for Computational Linguistics.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of NeurIPS*. Curran Associates, Inc.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for 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 CIKM*, page 2273–2276, New York, NY, USA. Association for Computing Machinery.

Tianshu Yu, Min Yang, and Xiaoyan Zhao. 2022. Dependency-aware prototype learning for few-shot relation classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2339–2345, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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. pages 3016–3025.

Peiyuan Zhang and Wei Lu. 2022. Better few-shot relation extraction with label prompt dropout. In *Proceedings of EMNLP*, pages 6996–7006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.