GRADUAL: Granularity-aware Dual Prototype Learning for Better Few-Shot Relation Extraction

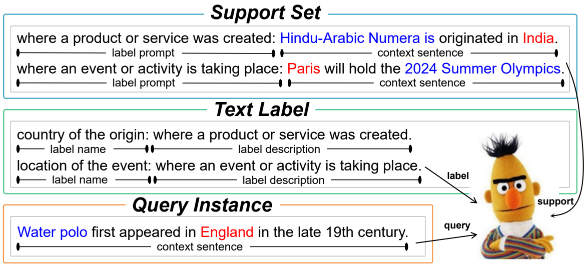
Zhiming Li[*] and **Yuchen Lyu**[*, †]
[1]School of Information Science and Engineering, Yanshan University, China
lizm@ysu.edu.cn, lyuyuchen.ysu@vip.163.com

Abstract

Recent studies have shown that fusing text labels and context sentences is an effective method for learning prototype representations in few-shot relation extraction. However, the *inconsistency of prototype representations* across different few-shot tasks persists due to different context sentences for the same relation, even with the integration of text labels into prototype representations. Conversely, the text label for each relation is unique and consistent, 1)which prompts us to propose a *dual prototype learning method*. Unlike previous methods that only construct support-based prototypes, we additionally construct label-based prototypes. Furthermore, we introduce a graph-based prototype adjustment module to construct topological information between support-based and label-based prototypes, thereby generating a more effective similarity measure through a simple linear combination. In addition, relations of different granularities have different distribution widths in the same semantic space, the *imbalanced distribution in the semantic space* leads to a lack of comparability among relations. To create a more discriminative semantic space, 2)we propose a *granularity-aware prototype learning method* that unifies the distribution width of relations, making relations of different granularities have similar distribution widths. Experimental results on two public benchmark datasets show that our proposed methods achieve state-of-the-art performance in few-shot relation classification. The source code can be accessed via the following link: https://github.com/ysulizm/GRADUAL.

1 Introduction

Relation Extraction (RE) is a fundamental task in Natural Language Processing (NLP) that aims to extract semantic relations between entities from natural language text (Ma et al., 2023). However, obtaining high-quality RE annotation data is time-consuming and labor-intensive. To alleviate the burden of manual annotation, some researchers have adopted Distant Supervision (DS) (Mintz et al., 2009) or Semi-Supervised Relation Extraction (SSRE) (Sun et al., 2011) to obtain annotation information. However, these methods can lead to inaccurate annotations and are not adaptable to situations with only a few labeled examples. To solve the problem of data scarcity, inspired by human cognitive mechanisms, researchers have proposed Few-Shot Learning (FSL) (Han et al., 2018). FSL is dedicated to leveraging learned prior knowledge to quickly generalize to new tasks that only contain a few labeled training samples.



where a product or service was created: Hindu-Arabic Numera is originated in India.

where an event or activity is taking place: Paris will hold the 2024 Summer Olympics.

country of the origin: where a product or service was created.

location of the event: where an event or activity is taking place.

Water polo first appeared in England in the late 19th century.

Figure 1: Data examples in a 2-way-1-shot scenario using the dual prototype learning method. The head and tail entities of the context sentence are represented in red and blue, respectively.

As a popular framework widely applied in few-shot tasks, Meta-Learning methods aim to learn the ability to learn quickly from experience and rapidly generalize to new subtasks (also known as meta-tasks) (Vinyals et al., 2016). Among them, Prototype Networks is a simple yet effective metric-based meta-learning method that aims to learn a metric space, and then classify query instances based on the distance between the query instances and the prototype representations (Snell et al., 2017).

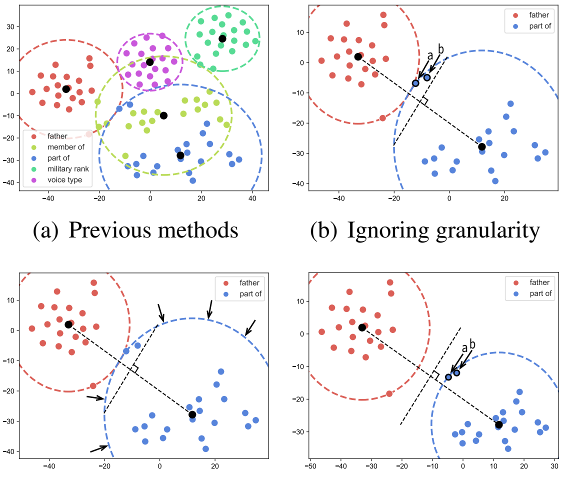
Despite the fact that Prototype Networks based

[*] Equal contribution.
[†] Correspondence author.

Findings of the Association for Computational Linguistics ACL 2024, pages 13566–13577
August 11-16, 2024 ©2024 Association for Computational Linguistics

on Pretrained Language Models (PLMs) have shown excellent performance in Few-shot Relation Extraction (FSRE), the prototype representation can be further enhanced by introducing external information, specifically, relation information, which can be achieved by integrating specific relation representations into the prototype representation (Baldini Soares et al., 2019; Peng et al., 2020). Therefore, many recent studies on introducing label information into Prototype Networks have achieved notable achievements. For example, Han et al. (2021a) proposed a hybrid prototype network that generates hybrid prototypes based on context sentences and relation descriptions to learn better prototype representations. Recently, Liu et al. (2022) found that even in the 1-shot setting, directly adding the embedding of relation descriptions and support sample representations to generate prototype representations can achieve good results. As the state-of-the-art model for implementing FSRE tasks, Zhang and Lu (2022) proposed a label prompt dropout method, which further enhances the prototype representation by directly concatenating the label and context sentence as input.

Although the above methods have achieved good performance, these works only consider integrating label information into prototype representations to learn better prototype representations, ignoring the fact that different context sentences of the same relation can lead to inconsistent prototype representations among different few-shot tasks. This inconsistency persists even when text labels are integrated into the prototype representation. Conversely, the text label for each relation is unique and consistent across different few-shot tasks, which prompts us to propose a novel method called *dual prototype learning method*. This method constructs support-based prototypes using context sentences, while using text labels to additionally construct label-based prototypes. Specifically, to enable the generated label-based prototypes to enhance the consistency of the support-based prototype in the same semantic space, we provide text labels and context sentences as input to the same encoder. A specific example of text labels and context sentences is shown in Figure 1. In addition, to enhance the consistency of support-based prototypes and better calculate the similarity between the prototype and the query instance, we employ a *graph-based prototype adjustment module* to construct the topology information between the support-based prototype and the label-based prototype. Based on the topological information, the similarity measure between the support-based prototype, the label-based prototype, and the query instance is generated through a simple linear combination. The new similarity measure can effectively calculate the similarity degree between the prototype and query instances by integrating both support-based and label-based prototype measures.



(a) Previous methods (b) Ignoring granularity

(c) Aware granularity (d) Our method

Figure 2: The t-SNE visualization of Prototype Networks in few-shot scenarios. The same color points belong to the same relation, and the dashed circle represents the distribution area of instances in the semantic space. (a) In the previous methods, relations of different granularities have different distribution widths in the same semantic space. (b) Because instances a and b are closer to the prototype of the "father" relation than the prototype of the "part of" relation, they are incorrectly classified as the "father" relation. (c) After the distribution widths of the "part of" relation and the "father" relation are unified, instances a and b will be correctly classified as the "part of" relation. (d) In our granularity-aware prototype learning method, relations of different granularities have similar distribution width in the same semantic space.

In addition, we emphasize a potential problem in previous research, where relations of different granularities in the same semantic space have different distribution widths. This imbalanced distribution leads to a lack of comparability between these relations, thereby influencing the classification results, as shown in Figure 2(a) and Figure 2(b). To address this problem, we propose a *granularity-aware prototype learning method* based on the *dual prototype learning method*. This method unifies the distribution widths of different relations in each few-shot task, thus constructing a new semantic space with stronger discrimination, as shown in Figure 2(c) and Figure 2(d). Specifically, this method first calculates the area formed by the label-based prototype, the support-based prototype, and the query instance. The area's size indicates the granularity of the relations, with larger areas denoting greater granularity. Then, the distribution width of relations in the same semantic space is adjusted according to their granularity. Notably, in our **Gra**nularity-aware **Dual** Prototype Learning Method (GRADUAL), relations of different granularities have similar distribution width in the same semantic space.

The main contributions of this paper are as follows:

- We propose a novel dual prototype learning method that includes a graph-based prototype adjustment module. This simple yet effective method outperforms previous approaches that required additional encoders or complex graph-based network structures to integrate text labels and context sentences.

- We introduce a granularity-aware prototype learning method based on the dual prototype learning method. This method unifies the distribution width sizes of different granularity relations for each few-shot task in the same semantic space, leading to the creation of a new semantic space with stronger discriminative power.

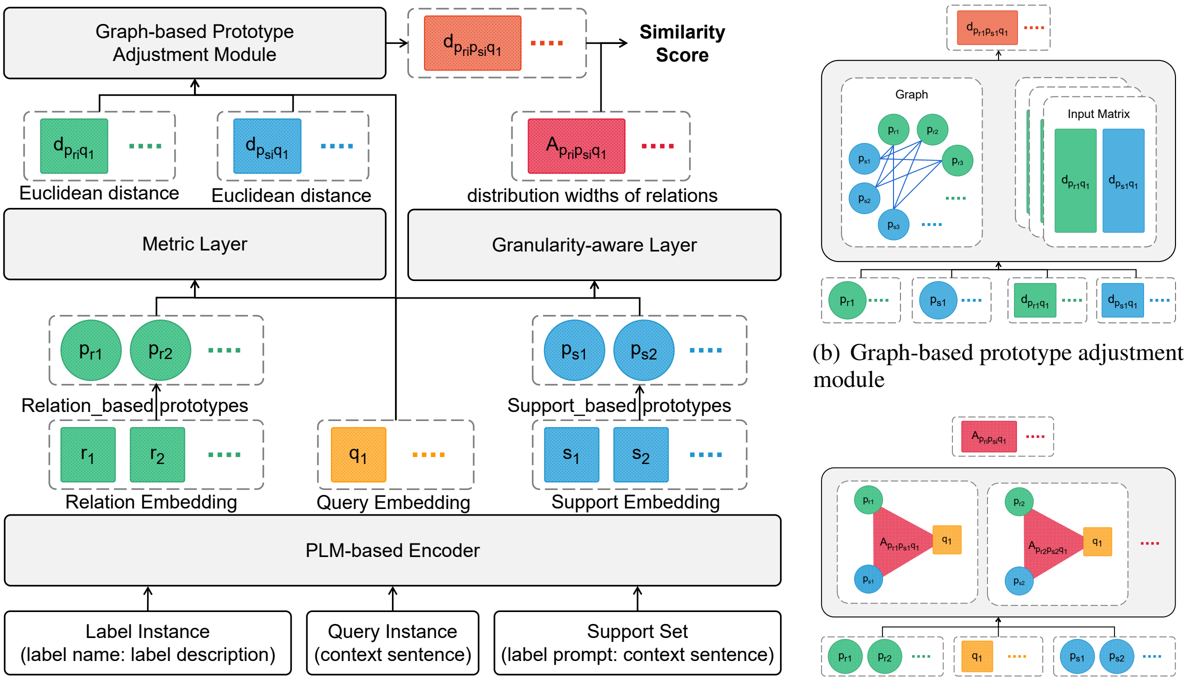
- We conduct extensive evaluations of GRADUAL under four few-shot settings on popular large-scale benchmark datasets for FSRE. The results demonstrate that our method outperforms existing state-of-the-art methods in FSRE tasks.

2 Related Work

Few-Shot Relation Extraction. Few-Shot Relation Extraction (FSRE) is an emerging field that extends Few-Shot Learning (FSL) to Relation Extraction (RE), enabling rapid adaptation to unseen relations with a small number of annotated instances (Garcia and Bruna, 2018). With the release of the large-scale benchmark dataset FewRel, FSRE has received increasing attention (Han et al., 2018; Gao et al., 2019). Current research on FSRE primarily falls into two categories: (1) **Methods based on Pretrained Language Models (PLMs)**, which further train PLMs on RE tasks to obtain better semantic space representations, such as MTB (Baldini Soares et al., 2019), CP (Peng et al., 2020), MapRE (Dong et al., 2021), LPD (Zhang and Lu, 2022), and others. (2) **Methods based on Metric Learning**, which aim to map the to-be-classified samples and the known classified samples to the same semantic space to compare their similarities (Li et al., 2024). Metric-based methods are non-parametric, easier to implement, and have shown strong performance in a series of few-shot tasks (Triantafillou et al., 2019). Therefore, they have been widely used in recent FSRE research, including Siamese Networks (Chen and He, 2021), Matching Networks (Han et al., 2020), Graph Neural Networks (Xie et al., 2020), and Prototype Networks (Snell et al., 2017). Among these, Prototype Networks has become the mainstream method for FSRE due to its efficiency.

Prototype Networks. Existing methods based on Prototype Networks often introduce external information (such as relation information) to enhance prototype representation, which is a key part of improving FSRE performance (Wu et al., 2024). Some methods add relation information to enhance prototype representation (Liu et al., 2022; Li et al., 2022). For example, Zhenzhen et al. (2022) employed a joint training method to learn the prototype encoder from relation definitions to enhance prototype representation. Other methods utilize relation information to highlight intra-class similarity and inter-class differences through **contrastive learning** (Wang et al., 2022) and **graph representation learning** (Yu et al., 2022). For instance, Li and Qian (2022) constructed a graph-based model generation framework to generate classification models according to the context sentence and text labels. Besides, some methods alleviate the distortion of prototype representation in the prototype network by storing prototype representations in memory through **continual learning** (Chen et al., 2023).

Summary. All the aforementioned methods have demonstrated the importance of high-quality prototype representation for FSRE. However, these methods also overlook the fact that relations of different granularities lead to imbalanced distribution in the semantic space. This imbalanced distribution reduces the model's performance and generalization ability. (1) To avoid introducing a large amount of irrelevant external information that can



(a) Model overview

(b) Graph-based prototype adjustment module

(c) Granularity-aware layer

Figure 3: An overview of our proposed GRADUAL.

mislead the model in learning false relevance, we only introduce relation information to enhance prototype representation in our dual prototype learning method. Furthermore, (2) to improve the semantic space's expressiveness, we propose a granularity-aware prototype learning method. This method transforms the imbalanced distribution of relations of different granularities in the semantic space into a similar distribution, thus creating a new semantic space with greater differentiation for FSRE.

3 Problem Formulation

We follow the typical few-shot task setting, namely the N-way-K-shot setup (Han et al., 2018). In each few-shot task, each instance c includes context sentence tokens $x = \{x_0, ..., x_m\}$, head entity $e^{head} = \{e^{head}_{start}, e^{head}_{end}\}$ and tail entity $e^{tail} = \{e^{tail}_{start}, e^{tail}_{end}\}$. The corresponding label $y = \{y^{text}, y^{num}\}$ to c contains a textual label y^{text} and a label index y^{num}, where $x_0 =$ [CLS] and $x_m =$ [SEP] represent the start and end positions of tokens, e^{head}_{start} and e^{tail}_{start} represent the start positions of entities, and e^{head}_{end} and e^{tail}_{end} represent the end positions of entities. Each N-way-K-shot task contains a support set $\mathcal{S} = \{s^i_j; i = 1, ..., N, j = 1, ..., K\}$ and a query set $\mathcal{Q} = \{q_i; i = 1, ..., M\}$, where S contains N relations, each relation has K different labeled instances, and $\mathcal{Q}$ contains M unlabeled instances. The aim of the few-shot task is to predict the correct label y for each query instance q in the query set.

Our model follows meta-learning setup during training (Vinyals et al., 2016), which consists of two stages: the meta-training stage and the meta-testing stage. The datasets for the meta-training and meta-testing stages are respectively referred to as the meta-training dataset $\mathcal{D}_{train}$ and the meta-testing dataset $\mathcal{D}_{test}$, with no overlapping relations between them. The meta-training dataset is divided into support instances s and query instances q to enable the model to gain transferable knowledge and the ability to "learn to learn". Unlike traditional model training methods, FSL, after integrating the concept of meta-learning, treats the meta-task as the training unit.

4 Methodology

4.1 Model Overview

The overview of our proposed model, GRADUAL, is shown in Figure 3. It mainly consists of 3 parts: (1) An encoder based on Pretrained Language Models (PLMs). (2) A graph-based prototype adjustment module in the dual prototype learning method. This module first constructs the topological information between the support-based prototype and the label-based prototype, and then generates a new similarity measure by combining different inputs (similarity measures between the support-based prototype and query instances, and between the label-based prototype and query instances) and the corresponding topological information through a simple linear combination. (3) A granularity-aware layer in the granularity-aware prototype learning method. The granularity-aware layer uses the area size between the support-based prototype, the label-based prototype, and the query instance to measure relations of different granularities in each few-shot task.

4.2 Encoder

Among various types of encoders used for relation extraction, encoders based on Pretrained Language Models (PLMs) have achieved notable results due to their impressive performance in capturing extensive general semantic knowledge. According to existing research, we adopt BERT$_{\text{BASE}}$, CP, and LPD as encoders to generate semantic information for all instances. For each support instance, to better utilize the implicit knowledge obtained by the encoder during training, we follow the settings of Zhang and Lu (2022), using the relation description and colon as label prompts to guide the output of the PLMs-based encoder. To construct a natural language sentence for each instance, we concatenate the label prompt with the context sentence. For example, in Figure 1, the context sentence "*Paris will hold the 2024 Summer Olympics*" concatenated with the corresponding label prompt becomes "*where an event or activity is taking place: Paris will hold the 2024 Summer Olympics*". For each query instance, we use the context sentence without any label prompts. For each label instance[1], following the settings of Liu et al. (2022), we concatenate the label name with the label description, separated by a colon. For example, in Figure 1, the label name "*location of the event*" concatenated with the corresponding label description becomes "*location of the event: where an event or activity is taking place*".

Subsequently, we introduce special tokens in each instance to distinguish head and tail entities, as well as to mark the start and end of context sentences. Specifically, we add [E1][/E1] and [E2][/E2] tokens at the positions of the head and tail entities, and insert [CLS] and [SEP] tokens at the start and end of instances. The support instance mentioned earlier is thus represented as "[CLS] *where an event or activity is taking place:* [E1] *Paris* [/E1] *will hold the* [E2] *2024 Summer Olympics* [/E2]. [SEP]". The label instance becomes "[CLS] *location of the event: where an event or activity is taking place.* [SEP]". Finally, the parsed instances are mapped into the semantic space through the encoder to obtain the corresponding text embeddings. We follow the settings of Li and Qian (2022). to obtain each instance representation. The representations of support and query instances are as follows:

$$\begin{aligned} s &= h^s_{e1} \oplus h^s_{e2} \\ q &= h^q_{e1} \oplus h^q_{e2} \end{aligned} \tag{1}$$

where s and q represent the representations of support and query instances respectively, h_{e1} and h_{e2} represent the embeddings of special tokens [E1] and [E2], and $\oplus$ represents concatenation operation. The representation of the label instance as follows:

$$r = h_0 \oplus \frac{\sum_{i=1}^{n} h_i}{n} \tag{2}$$

where h_0 represents the embedding of the special token [CLS], and n represents the number of tokens in the text embedding.

4.3 Dual Prototype Learning Method

Obtaining high-quality prototype representations is crucial to improve the performance of FSRE. Therefore, existing methods consider integrating label information into prototype representations to learn more effective prototype representations. However, different context sentences of the same relation can lead to inconsistent prototype representations across different few-shot tasks.To address this issue, we propose a dual prototype learning method that alleviates this inconsistency by generating consistent label-based prototypes. To more effectively calculate the similarity between prototypes and query instances, we introduce a graph-based prototype adjustment module. This module first constructs topological information between support-based prototypes and label-based prototypes, and then generates a new similarity measure based on topological information by integrating both support-based and label-based prototype measures.

For each N-way-K-shot task, the support-based prototype p_s for each relation is obtained by aver-

[1] In this study, the terms text label and label instance are different. Text labels contain label names and label descriptions, and label instances are concatenated by label names and label descriptions.

aging the representations of K support instances in that relation, as described by Snell et al. (2017). The support-based prototype p_s as follows:

$$p_s = \frac{1}{K} \sum_{i=1}^{K} s_i \tag{3}$$

Meanwhile, The label instance representation for each relation is directly used as the label-based prototype p_r in that relation. The label-based prototype p_r as follows:

$$p_r = r \tag{4}$$

Subsequently, we use the Euclidean distance as our similarity measure function to calculate the degree of similarity between various instances. For each N-way-K-shot task, the Euclidean distance between the query instance and the support-based prototype as follows:

$$d_{p_s q} = d(p_s, q) \tag{5}$$

where $d(\cdot)$ represents the Euclidean distance. The Euclidean distance between the query instance and the label-based prototype as follows:

$$d_{p_r q} = d(p_r, q) \tag{6}$$

In the graph-based prototype adjustment module, to construct topological information between the support-based prototype and the label-based prototype, we consider the support-based prototype and all label-based prototypes as nodes in the graph and then establish edges between the nodes. Specifically, for each N-way-K-shot task, we first calculate the Euclidean distance between the support-based prototype and all label-based prototypes as follows:

$$d_{p_s p_r} = d(p_s, p_r) \tag{7}$$

and we can get the smallest Euclidean distance as follows:

$$d_{p_s min} = \arg \min_{i \in \{1,...N\}} d(p_s p_{r_i}) \tag{8}$$

We then calculate the ratio of $d_{p_s min}$ to $d_{p_s p_r}$ as λ, which represents the edge between nodes, as follows:

$$\lambda = \frac{d_{p_s min}}{d_{p_s p_r}} \tag{9}$$

In our graph, the edge represents the ratio of the similarity measure between two nodes of the support-based prototype and the label-based prototype that need to be reduced to make them belong to the same label. Finally, according to the topological information in the graph, we generate a new similarity measure by integrating both support-based and label-based prototype measures through a simple linear combination:

$$d_{p_s p_r q} = d_{p_s q} + \lambda d_{p_r q} \tag{10}$$

4.4 Granularity-aware Prototype Learning Method

To alleviate the imbalanced distribution of different granularity relations in the same semantic space, we unify the distribution width of different relations, thereby constructing a new semantic space with more discrimination. Specifically, in the granularity-aware layer, we first calculate the area $\mathcal{A}$ formed by the label-based prototype, the support-based prototype, and the query instance as follows:

$$\begin{gathered} \mathcal{A} = \sqrt{p(p - d_{p_s q})(p - d_{p_r q})(p - d_{p_r p_s})} \\ p = \frac{d_{p_s q} + d_{p_r q} + d_{p_r p_s}}{2} \end{gathered} \tag{11}$$

We then unify the distribution width of different relations in each few-shot task, so that different granularity relations have similar distribution widths. Therefore, the new similarity measure $d_{p_s p_r q}$ obtained by our dual prototype learning method becomes $d^{\mathcal{A}}_{p_s p_r q}$, as follows:

$$d^{\mathcal{A}}_{p_s p_r q} = \frac{d_{p_s p_r q}}{\mathcal{A}} \tag{12}$$

Subsequently, we use $d^{\mathcal{A}}_{p_s p_r q}$ as the logit in the cross-entropy loss as follows:

$$\mathcal{L} = -\sum_{i=1}^{N} \log \frac{\exp\left(d^{\mathcal{A}}_{p_{s_i} p_{r_i} q}\right)}{\sum_{j=1}^{N} \exp\left(d^{\mathcal{A}}_{p_{s_j} p_{r_j} q}\right)} \tag{13}$$

We calculate the gradient by using the cross-entropy loss and process the gradient update on the encoder to make our model reach the optimal point.

5 Experimental Setup

5.1 Datasets

Our GRADUAL is evaluated on the FewRel 1.0 (Han et al., 2018) and the domain adaptation part of FewRel 2.0 (Gao et al., 2019) datasets. FewRel 1.0

consists of 100 relations, with each relation containing 700 labeled instances. Our experiments follow the split used in the official benchmark test, where the dataset is divided into 64 base relations for training, 16 relations for validation, and 20 new relations for testing. Unlike FewRel 1.0, which conducts training and testing in the same Wikipedia domain, FewRel 2.0, which has domain adaptation capabilities, conducting training in the Wikipedia domain and testing in the biomedical domain to evaluate the model's domain transferability. Since the labels of the test sets for FewRel 1.0 and FewRel 2.0 are not publicly accessible, we submit the model's predicted results to CodaLab to obtain the accuracy of the test set.

5.2 Compared Methods

We compare GRADUAL with 17 existing baselines. According to the type of encoder, we divide these baselines into three groups: standard BERT-based baselines, BERT-based baselines that introduce external information, and baselines based on pre-trained language models (PLMs).

In the group of standard BERT-based baselines, 1) **Proto-BERT** (Han et al., 2018) is a prototype network that uses BERT as the encoder. 2) **MAML** (Finn et al., 2017) is an optimization-based meta-learning method. 3) **GNN** (Satorras and Estrach, 2018) is a graph neural network-based meta-learning method. 4) **BERT-PAIR** (Gao et al., 2019) is a sequence classification model that measures the similarity between two instances.

In the group of BERT-based baselines that introduce external information, 5) **REGRAB** (Qu et al., 2020) and 6) **ConceptFERE** (Yang et al., 2021) model different relations by leveraging a global relation graph and the inherent concepts of entities, respectively. 7) **CTEG** (Wang et al., 2020) is a model that decouples high co-occurrence relations with various external information. 8) **TD-Proto** (Yang et al., 2020) enhances the prototype network by introducing text descriptions of Wikidata. 9) **HCRP** (Han et al., 2021a) distinguishes task complexity by introducing relation descriptions. 10) **DRK** (Wang et al., 2022) uses a rule-based knowledge method to alleviate prediction confusion. 11) **SimpleFSRE** (Liu et al., 2022) directly integrates relation descriptions into prototype representations. 12) **GM_GEN** (Li and Qian, 2022) constructs topological information to generate specific classification models.

In the group of PLMs-based baselines, 13) **MTB** (Baldini Soares et al., 2019) is a pre-trained model for matching blank tasks based on BERT$_{\text{LARGE}}$. 14) **CP** (Peng et al., 2020) proposes an entity masking pre-trained framework based on contrastive learning. 15) **LDUR** (Han et al., 2021b) develops a supervised contrastive pre-trained method for learning discriminative representations. 16) **MapRE** (Dong et al., 2021) additionally uses a relation encoder during pre-training to consider relation information. 17) **LPD** (Zhang and Lu, 2022) proposes a label prompt dropout method by directly concatenating labels and context sentences. In addition, HCRP (CP), SimpleFSRE (CP), and GM-GEN (CP) are standard CP-based baselines. To make a fair comparison with these baselines, our GRADUAL provides experimental results with BERT$_{\text{BASE}}$, CP, and LPD as encoders.

5.3 Implementation Details

We conduct our experiments on servers equipped with 24GB NVIDIA RTX 3090. We use uncased BERT$_{\text{BASE}}$, CP, and LPD as instance encoders for fair comparison with other models, and optimize our GRADUAL with AdamW (Loshchilov and Hutter, 2019). BERT$_{\text{BASE}}$, CP, and LPD are all composed of 12 layers of transformer modules; in addition, CP is further pre-trained through contrastive learning, while LPD introduces a label prompt dropout method and combines it with contrastive learning for further pre-training. Table 5 shows the detailed hyperparameters. To obtain the accuracy of the test set, we follow the official evaluation settings and submit the predicted results to the FewRel leaderboard. In Table 1 and Table 2, we report the average accuracy and standard deviation of 5 runs with different random seeds.

6 Results and Analyses

6.1 Main Results

The comparison results of FewRel 1.0 and 2.0 are shown in Table 1 and Table 2, respectively. The detailed information of the baseline methods compared is provided in Chapter 5.2. In addition, we directly utilize the settings and results from GM-GEN and LPD for FewRel 2.0. Our findings from the comparison results are as follows:

(1) **GRADUAL effectively addresses the FSRE problem.** As shown in Table 1 and Table 2, our proposed GRADUAL significantly outperforms all methods that use the same encoder. Moreover, we find that among the compared baseline

Model	5-way-1-shot		5-way-5-shot		10-way-1-shot		10-way-5-shot	
	val	test	val	test	val	test	val	test
MAML♣ (Finn et al., 2017)	82.93	89.70	86.21	83.55	73.20	83.17	86.06	88.51
GNN♣ (Satorras and Estrach, 2018)	-	75.66	-	89.06	-	70.08	-	76.93
Proto-BERT♣ (Han et al., 2018)	82.92	80.68	91.32	89.60	73.24	71.48	83.68	82.89
BERT-PAIR♠ (Gao et al., 2019)	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
CTEG (Wang et al., 2020)	84.72	88.11	92.52	95.25	76.01	81.29	84.89	91.33
TD-Proto (Yang et al., 2020)	-	84.76$_{\pm 0.20}$	-	92.38$_{\pm 0.11}$	-	74.32$_{\pm 0.12}$	-	85.92$_{\pm 0.06}$
ConceptFERE (Yang et al., 2021)	-	89.21	-	90.34	-	75.72	-	81.82
HCRP (Han et al., 2021a)	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
SimpleFSRE (Liu et al., 2022)	91.29	94.42	94.05	96.37	86.09	90.73	89.68	93.47
GM_GEN (Li and Qian, 2022)	92.65	94.89	95.62	96.96	86.81	91.23	91.27	94.30
GRADUAL	**92.54**$_{\pm 0.51}$	**95.55**$_{\pm 0.28}$	**96.28**$_{\pm 0.09}$	**97.03**$_{\pm 0.09}$	**87.46**$_{\pm 0.42}$	**91.70**$_{\pm 0.35}$	**92.40**$_{\pm 0.29}$	**94.33**$_{\pm 0.31}$
MTB♡ (Baldini 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
LDUR (Han et al., 2021b)	87.21	90.40	94.86	96.95	80.34	84.68	91.36	94.15
MapRE (Dong et al., 2021)	-	95.73	-	97.84	-	93.18	-	95.64
HCRP (CP) (Han et al., 2021a)	94.10	96.42	96.05	97.96	89.13	93.97	93.10	96.46
SimpleFSRE (CP) (Liu et al., 2022)	96.21	96.63	97.07	97.93	93.38	94.94	95.11	96.39
GM_GEN (CP) (Li and Qian, 2022)	96.97	97.03	98.32	98.34	93.97	94.99	96.58	96.91
GRADUAL (CP)	**97.47**$_{\pm 0.13}$	**97.64**$_{\pm 0.09}$	**98.59**$_{\pm 0.04}$	**98.66**$_{\pm 0.02}$	**95.30**$_{\pm 0.16}$	**95.89**$_{\pm 0.22}$	**97.12**$_{\pm 0.04}$	**97.35**$_{\pm 0.06}$
LPD_filtered (Zhang and Lu, 2022)	93.51$_{\pm 0.7}$	95.12$_{\pm 0.2}$	94.33$_{\pm 0.7}$	95.79$_{\pm 0.1}$	87.77$_{\pm 1.1}$	90.73$_{\pm 0.2}$	89.19$_{\pm 1.3}$	92.15$_{\pm 0.3}$
GRADUAL (LPD_filtered)	**94.62**$_{\pm 0.34}$	**96.47**$_{\pm 0.21}$	**96.22**$_{\pm 0.22}$	**97.26**$_{\pm 0.12}$	**89.82**$_{\pm 0.33}$	**92.63**$_{\pm 0.39}$	**92.09**$_{\pm 0.11}$	**94.28**$_{\pm 0.48}$
LPD (Zhang and Lu, 2022)	97.76$_{\pm 0.1}$	98.17$_{\pm 0.0}$	97.75$_{\pm 0.2}$	98.29$_{\pm 0.2}$	96.21$_{\pm 0.2}$	96.66$_{\pm 0.0}$	96.28$_{\pm 0.1}$	96.75$_{\pm 0.2}$
GRADUAL (LPD)	**98.44**$_{\pm 0.10}$	**98.71**$_{\pm 0.07}$	**98.64**$_{\pm 0.05}$	**98.84**$_{\pm 0.04}$	**96.99**$_{\pm 0.12}$	**97.77**$_{\pm 0.06}$	**97.06**$_{\pm 0.05}$	**97.79**$_{\pm 0.06}$

Table 1: Accuracy (%) of FSRE on FewRel 1.0 validation/test set. ♠ represents results from the public FewRel leaderboard, ♣ represents results reported by Qu et al. (2020), and ♡ represents results reported by Peng et al. (2020). Our GRADUAL provides experimental results using BERT$_{\mathrm{BASE}}$, CP, and LPD as encoders, respectively. In particular, LPD_filtered represents LPD pretrained on the Wikipedia (filtered) dataset (Zhang and Lu, 2022). Detailed information can be found in Chapter 5.2. The experimental data from our model are represented in **bold**.

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
Proto-ADV	42.21	58.71	28.91	44.35
BERT-PAIR	67.41	78.57	54.89	66.85
HCRP	76.34	83.03	63.77	72.94
GM-GEN	76.67	91.28	64.19	84.84
GRADUAL	**81.71**$_{\pm 0.91}$	**91.49**$_{\pm 0.08}$	**72.59**$_{\pm 0.37}$	**83.72**$_{\pm 0.56}$
CP	79.70	84.90	68.10	79.80
GRADUAL (CP)	**84.99**$_{\pm 0.42}$	**92.48**$_{\pm 0.18}$	**75.02**$_{\pm 0.28}$	**86.19**$_{\pm 0.46}$
LPD_filtered	83.41$_{\pm 0.5}$	90.00$_{\pm 0.3}$	73.28$_{\pm 0.8}$	81.80$_{\pm 0.9}$
GRADUAL (LPD_filtered)	**85.76**$_{\pm 1.23}$	**92.11**$_{\pm 0.27}$	**75.82**$_{\pm 0.40}$	**84.62**$_{\pm 0.29}$
LPD	82.81$_{\pm 0.5}$	88.98$_{\pm 1.4}$	70.51$_{\pm 1.5}$	78.76$_{\pm 1.6}$
GRADUAL (LPD)	**84.98**$_{\pm 0.98}$	**91.63**$_{\pm 0.30}$	**76.13**$_{\pm 0.98}$	**84.52**$_{\pm 0.43}$

Table 2: Accuracy (%) of FSRE on FewRel 2.0 test set.

methods, those based on LPD perform better than those using CP and BERT, and methods based on CP also perform better than those based on BERT. However, the accuracy of BERT-based GRADUAL can be higher than some of the CP-based baseline methods, and the accuracy of CP-based GRADUAL can be higher than some of the LPD-based baseline methods. This shows that compared to these advanced baseline methods, our GRADUAL can solve the FSRE problem more effectively.

(2) **GRADUAL makes better use of external information.** As shown in Table 1, our proposed GRADUAL outperforms all BERT-based baselines that introduce external information. This is because our dual prototype learning method uses label information to generate consistent prototype representations for each relation, and combines support-based prototype measurements and label-based prototype measurements into a new similarity measurement by constructing topological information between prototypes, which helps the model to obtain higher quality prototype representations.

(3) **GRADUAL has robust generalization capabilities.** As shown in Table 2, compared to other baseline methods, GRADUAL demonstrates better adaptability to unknown data from different domains. This superior adaptability is due to our introduction of a granularity-aware prototype learning method based on the dual prototype learning method. By transforming the imbalanced distribution of relations of different granularities in the same semantic space into a similar distribution, we construct a new semantic space with greater discrimination for FSRE. This new semantic space shows excellent adaptability to unknown data of relations of different granularities from different domains.

To further demonstrate the superiority of our model, an **anomaly analysis** is conducted on specific values of Table 2. Specifically, the accuracy of

GM_GEN using the BERT encoder is 84.84, which is lower than the accuracy of our CP-based GRADUAL at 86.19 but higher than our BERT-based GRADUAL at 83.72. Notably, among the four paradigms, the 10-way-1-shot and 5-way-1-shot paradigms are the most challenging, and our model achieves the highest accuracy in both paradigms using the same encoder. Therefore, when using the same BERT encoder in the FewRel 2.0 dataset, our model's performance is indeed superior to GM_GEN.

Model	5-way-1-shot		5-way-5-shot		10-way-1-shot		10-way-5-shot	
	1.0	2.0	1.0	2.0	1.0	2.0	1.0	2.0
DUAL_Base	94.86	75.43	96.98	90.46	91.10	62.87	93.58	84.26
DUAL	94.87	78.12	96.97	91.92	91.09	65.62	93.58	86.48
GRADUAL_Base	95.17	79.59	97.28	90.06	91.75	67.40	94.65	79.35
GRADUAL	95.57	81.71	97.30	91.85	92.01	70.73	94.97	83.97
DUAL_Base (CP)	96.46	79.02	98.25	92.10	92.87	66.75	96.84	85.62
DUAL (CP)	96.48	82.18	98.29	92.52	92.88	73.07	96.75	86.27
GRADUAL_Base (CP)	97.74	83.85	98.68	92.47	95.78	71.82	97.34	85.69
GRADUAL (CP)	97.68	85.24	98.71	92.70	96.00	75.16	97.42	86.38

Table 3: Ablation results in accuracy (%) on FewRel 1.0/2.0 test set.

6.2 Ablation Study

We conduct several ablation experiments on the Fewrel dataset using GRADUAL to examine the relative contributions of different components in the model shown in Table 3. Here, DUAL represents the dual prototype learning method, GRADUAL represents a granularity-aware prototype learning method built on DUAL, and DUAL_Base and GRADUAL_Base represent methods without the graph-based prototype adjustment module.

(1) **Graph-based prototype adjustment module.** From Table 3, we find that after incorporating the graph-based prototype adjustment module into the method, the accuracy of the method significantly improved. This suggests that the graph-based prototype adjustment module in the dual prototype learning method can indeed enhance the prototype representation, thereby making the generated similarity measures more appropriate for FSRE.

(2) **Granularity-aware layer.** We also find from Table 3 that the accuracy of GRADUAL increased relative to DUAL, demonstrating the effectiveness of the granularity-aware layer in GRADUAL. However, when BERT is used as the encoder in the experiments on the FewRel 2.0 dataset, we observe a decrease in accuracy in some results. The decrease in accuracy is due to the oversimplification of the granularity measurement method, indicating that there is potential for improvement in our granularity measurement method.

7 Conclusions

In this work, we introduce GRADUAL, an innovative method designed for Few-Shot Relation Extraction (FSRE). GRADUAL enhances the performance and generalization capability of the model by generating more consistent prototype representations for each relation and constructing a more discriminative semantic space for different relations. Firstly, we first combine the metrics of support-based prototypes and label-based prototypes into a new similarity metric by constructing topological information between prototypes, which helps the model to get high-quality prototype representations. Secondly, to enhance the expressive power of the semantic space, we measure the granularity of different relations by the area size between the prototype and the query instance. By transforming the imbalanced distribution of different granularity relations in the same semantic space into a similar distribution, we construct a more discriminative semantic space. Extensive experiments have demonstrated that our GRADUAL significantly outperforms previous works. In future research, we plan to apply GRADUAL to more information extraction tasks across various domains and examine its robustness.

8 Limitations

Despite GRADUAL achieves new state-of-the-art performance in FSRE, it still has several limitations. Firstly, we combine two prototype representations using simple topological information, but more advanced combination methods could yield higher quality prototype representations. Secondly, our method for measuring granularity is relatively simple. Although it is effective, we believe that more advanced methods for measuring granularity could further improve the model performance. Lastly, we have not yet explored the effectiveness of GRADUAL in other text classification tasks, such as intent classification. We believe these areas present promising directions for future research.

Acknowledgments

This work has been partially supported by the grants from S&T Program of Hebei No.20375001D; Project of Hebei Key Laboratory of Software Engineering, No.22567637H, No.ZZYB2302.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753.

Xiudi Chen, Hui Wu, and Xiaodong Shi. 2023. Consistent prototype learning for few-shot continual relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7409–7422, Toronto, Canada. 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.

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*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

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 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.

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

Jiale Han, Bo Cheng, and Wei Lu. 2021a. 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.

Jiale Han, Bo Cheng, and Guoshun Nan. 2021b. Learning discriminative and unbiased representations for few-shot relation extraction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 638–648, New York, NY, USA. Association for Computing Machinery.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Jie Zhou, and Maosong Sun. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 745–758, Suzhou, China. 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.

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.

Wanli Li and Tieyun Qian. 2022. Graph-based model generation for few-shot relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 62–71, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xuewei Li, Chao Liu, Jian Yu, Tianyi Xu, Mankun Zhao, Hongwei Liu, Mei Yu, and Ruiguo Yu. 2022. Prototypical attention network for few-shot relation classification with entity-aware embedding module. *Applied Intelligence*, 53(9):10978–10994.

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 the Association for Computational Linguistics: ACL 2022*, 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*.

Ruotian Ma, Zhang Lin, Xuanting Chen, Xin Zhou, Junzhe Wang, Tao Gui, Qi Zhang, Xiang Gao, and Yun Wen Chen. 2023. Coarse-to-fine few-shot learning for named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4115–4129, Toronto, Canada. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel 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, Suntec, Singapore. Association for Computational Linguistics.

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 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*, volume 119 of *Proceedings of Machine Learning Research*, pages 7867–7876. PMLR.

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

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ang Sun, Ralph Grishman, and Satoshi Sekine. 2011. Semi-supervised relation extraction with large-scale word clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 521–529, Portland, Oregon, USA. Association for Computational Linguistics.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. 2019. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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.

Hui Wu, Yuting He, Yidong Chen, Yu Bai, and Xiaodong Shi. 2024. Improving few-shot relation extraction through semantics-guided learning. *Neural Networks*, 169:453–461.

Yuxiang Xie, Hua Xu, Jiaoe Li, Congcong Yang, and Kai Gao. 2020. Heterogeneous graph neural networks for noisy few-shot relation classification. *Knowledge-Based Systems*, 194:105548.

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

Tianyuan Yu, Sen He, Yi-Zhe Song, and Tao Xiang. 2022. Hybrid graph neural networks for few-shot learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):3179–3187.

Peiyuan Zhang and Wei Lu. 2022. Better few-shot relation extraction with label prompt dropout. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6996–7006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

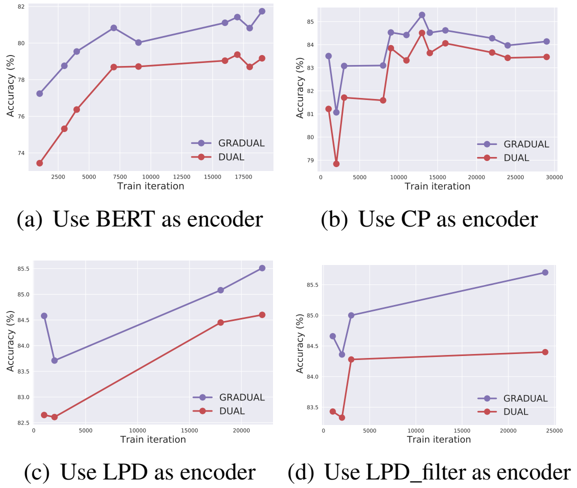
Li Zhenzhen, Yuyang Zhang, Jian-Yun Nie, and Dongsheng Li. 2022. Improving few-shot relation classification by prototypical representation learning with definition text. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 454–464, Seattle, United States. Association for Computational Linguistics.

A Analysis of the Effectiveness of the granularity-aware layer

Model	**BERT**	**CP**	**LPD**	**LPD_filter**
DUAL	2.10	1.62	1.10	0.56
GRADUAL	1.45	1.09	0.77	0.58

Table 4: Standard deviations of local optimal points generated during the training process in the 5-way-1-shot task.

From Figure 4, we can find that the accuracy of GRADUAL is consistently higher than that of DUAL, indicating that the granularity-aware layer



(a) Use BERT as encoder (b) Use CP as encoder

(c) Use LPD as encoder (d) Use LPD_filter as encoder

Figure 4: Accuracy of DUAL and GRADUAL in the training process of 5-way-1-shot task, based on different encoders. DUAL represents GRADUAL without the granularity-aware layer.

can effectively help the model construct a more discriminative semantic space, thereby improving its performance in FSRE. In addition, due to the superior performance of the LPD encoder over BERT and CP encoders, the method using LPD generates fewer local optimal points during the training process. From Table 4, we find that the standard deviation of GRADUAL is generally lower than that of DUAL, suggesting that the granularity-aware layer can not only improve the performance of the model, but also further enhance the stability of the model.

B Analysis of the Effectiveness of Graph-Based Prototype Adjustment Module

To help us study the importance of the graph-based prototype adjustment module in GRADUAL, we introduce a hyperparameter α into λ, hence the original Formula 9 now becomes as follows:

$$\lambda = (\frac{d_{p_s min}}{d_{p_s p_r}})^{\alpha} \tag{14}$$

From Figure 5, we can observe that the variation of λ indeed significantly impacts the accuracy of GRADUAL. Since λ mainly affects the label-based prototype, this shows the necessity of the label-based prototype, and also shows that our proposed dual prototype learning method can obtain higher-quality prototype representations.

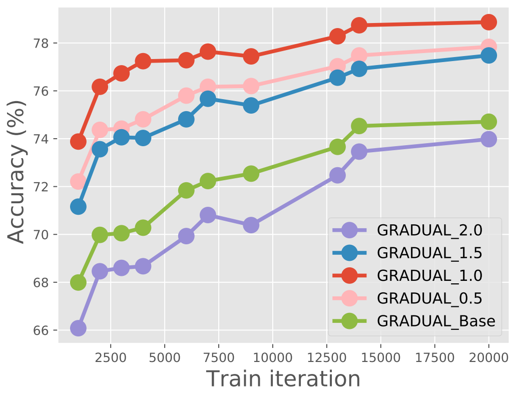


Figure 5: Accuracy of models based on the BERT encoder during the training process in the 5-way-1-shot task. Here, GRADUAL_Base represents GRADUAL without the graph-based prototype adjustment module, while GRADUAL_2.0, GRADUAL_1.5, GRADUAL_1.0, and GRADUAL_0.5 represent GRADUAL with α set to 2.0, 1.5, 1.0, and 0.5, respectively.

C Hyperparameters of GRADUAL

Dataset	Parameter	Value
Fewrel 1.0	encoder	BERT$_{\text{BASE}}$/CP/LPD
	random seed	41/42/43/44/45
	hidden size	768
	max length	128
	learning rate	$5e-6$
	batch size	4
	train iteration	20000
	val iteration	10000
Fewrel 2.0	encoder	BERT$_{\text{BASE}}$/CP/LPD
	random seed	41/42/43/44/45
	hidden size	768
	max length	128
	learning rate	$2e-5$
	batch size	4
	train iteration	20000

Table 5: Hyperparameters of GRADUAL.