# Graph-based Model Generation for Few-Shot Relation Extraction

**Wanli Li** and **Tieyun Qian***
School of Computer Science, Wuhan University, China
{wanli.li, qty}@whu.edu.cn

## Abstract

Few-shot relation extraction (FSRE) has been a challenging problem since it only has a handful of training instances. Existing models follow a 'one-for-all' scheme where one general large model performs all individual *N*-way-*K*-shot tasks in FSRE, which prevents the model from achieving the optimal point on each task. In view of this, we propose a model generation framework that consists of *one general model* for all tasks and *many tiny task-specific models* for each individual task. The general model *generates and passes the universal knowledge to* the tiny models which will be further fine-tuned when performing specific tasks. In this way, we decouple the complexity of the entire task space from that of all individual tasks while absorbing the universal knowledge. Extensive experimental results on two public datasets demonstrate that our framework reaches a new state-of-the-art performance for FRSE tasks. Our code is available at: https://github.com/NLPWM-WHU/GM_GEN.

## 1 Introduction

Relation extraction (RE) aims to detect the implied relations between/among entities mentioned in sentences. It plays a significant role in natural language processing (NLP). Massive unstructured text can be transformed into structured factual knowledge using the RE technique. However, training RE models requires sufficient human-annotated data. The models' performance usually drops dramatically without enough training data. However, in realistic scenarios, the acquisition of high-quality annotated data on RE is time-consuming and labor-intensive (Han et al., 2018; Gao et al., 2019b). To alleviate the labeled data scarcity problem, some researchers employ distant supervision (DS) (Reiplinger et al., 2014) or semi-supervised relation extraction (SSRE) (Lin et al., 2019; Li et al., 2022) techniques to improve the model performance.

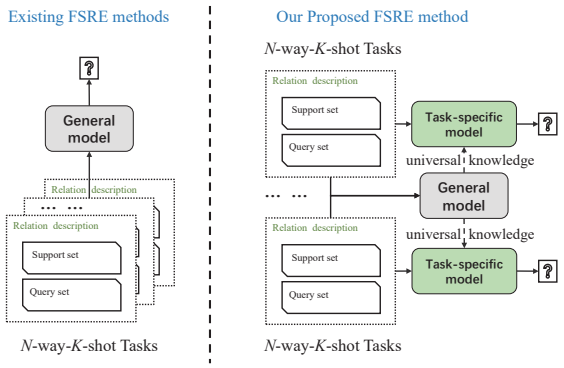



Figure 1: An illustration of the difference between existing methods and our proposed framework.

DS and SSRE obtain annotation information with the help of knowledge base and model prediction results, respectively. DS assumes that if two entities belong to a relation in KB, all sentences mentioning these entities indicate the relation class in KB. These methods inevitably generate false-positive and false-negative samples. SSRE methods require many unlabeled data and iteratively train a new model using the results on unlabeled data predicted by the trained model and they cannot adapt to the situation with only a few labeled examples either. That is to say, it is still an urgent problem to train a successful model under an extreme data scarcity condition where entire categories are introduced with just one or few examples.

A more challenging, yet practical extension, is the few-shot relation extraction (FSRE) task which handles the RE tasks with a handful of training instances. To better understand FSRE, we illustrate the *N*-way-*K*-shot scenario in FSRE where *N* and *K* respectively represent the category number and the samples per category. We provide a *10*-way-*5*-shot example here. Given *the relation description*, 10*5 labeled sentences as *a support set*, and 10 unlabeled sentences as *a query set*, the FSRE model will predict the relation between two entities mentioned

*Corresponding author



*Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 62 - 71
December 7-11, 2022 

in each sentence of the query set with the help of the support set and possible external knowledge.

Existing FSRE methods often employ metric-based learning techniques to improve FSRE performance, which trains a function $e_\phi(\cdot)$ mapping all samples into one same embedding space and then using some distance functions in the embedding space to label query samples. The intuition is that the distance between the query and the instances with the same relation is smaller than those with different relations. More recent FSRE methods integrate various information, such as entity description (Yang et al., 2020), relation description (Han et al., 2021a; Liu et al., 2022), to enhance the prototypical network. The aforementioned methods try to build a general embedding function learned from all training instances and utilize distance metrics to label instances. However, a general model cannot adapt to the single $N$-way-$K$-shot task because there exists a discrepancy between the entire task space and each individual task. The general model can only reach the optimal point in the whole task space instead of each task space. Consequently, such a 'one-for-all' learning scheme (Zhmoginov et al., 2022), i.e. one model solves all individual $N$-way-$K$-shot tasks, seriously limits the performance of FSRE models.

In this paper, we propose a novel model generation framework to overcome the limitation of the 'one-for-all' learning scheme in FSRE. The basic idea is to train one general model for all tasks and many tiny models for each individual $N$-way-$K$-shot task. As shown in Figure 1, we pre-train a general model to generate tiny task-specific models. During the testing phase, the tiny models can be fine-tuned according to the task-specific support samples and labels without affecting the knowledge from the general model. Specifically, we believe that the topology information (graph) between the samples and the label representation should also be modeled in addition to the semantic features called attributes of the sample. Therefore, we first employ a graph-based model generation module to combine the topology information with the attributes of instances and the relation descriptions. Then, the graph-based model generates many tiny classification models which will be fine-tuned and infer on different few-shot tasks. The separation of the general model and task-specific models successfully decouples the complexity of the entire task space from that of all individual tasks yet absorbing the universal knowledge.

The contributions of this paper are as follows:

- We re-frame FSRE tasks within a new model generation paradigm. To the best of our knowledge, we are the first to introduce this technique into FSRE.
- We develop a graph-based model generation module that can integrate both topology and attribute information and generate the universal knowledge across all tasks.
- Extensive experimental results demonstrate that our framework can achieve a new state-of-the-art performance for FSRE tasks.

## 2 Related Work

### 2.1 Few-shot learning

As a branch of machine learning, few-shot learning (FSL) attracts much attention due to its generalization ability to new domains (Koch et al., 2015). Existing FSL methods can be roughly classified into metric-based learning, optimization-based learning, and weight modulation methods. Metric-based learning methods (Snell et al., 2017a,b) learn a function $e_\phi(\cdot)$ mapping all samples into the same vector space and then use some nearest neighbor algorithms to label those query samples. Many optimization-based learning methods (Finn et al., 2017; Nichol et al., 2018; Rusu et al., 2019; Antoniou et al., 2019) argue that the gradient produced by a single task is not globally optimal. Therefore they fine-tune the embedding $e_\phi(\cdot)$ by performing additional SGD updates on all parameters of the model. Weight modulation methods (Li et al., 2019), can learn transferable prior knowledge across tasks and produce network parameters for similar unseen tasks with training samples. One latest method (Zhmoginov et al., 2022) directly employs a transformer-based model to generate weights of a convolutional neural network. In our work, we design a graph-based model generation approach that is more suitable for FSRE tasks.

### 2.2 Few-shot relation extraction

Few-shot relation extraction (FSRE) is a branch of relation extraction task, aiming at predicting semantic relations between head and tail entities mentioned in a given instance with a few labeled instances. Han et al. (2018) first propose FewRel, a large-scale dataset to explore few-shot learning

in relation extraction. Late research in FSRE has gradually developed into two categories. The first line of FSRE research only utilizes the text data and the provided relation description information without any external information. For example, Proto-HATT (Gao et al., 2019a) employs an attention mechanism to promote the prototypical network. HCRP (Han et al., 2021a) distinguishes difficult tasks from easy ones by introducing relation description information. SimpleFSRE (Liu et al., 2022) finds that directly adding relation description embedding and support sample representation can achieve good results even in a 1-shot setting.

Another line of FSRE research introduces external information to compensate for the extremely limited information. For instance, REGRAG (Qu et al., 2020) constructs a global relation graph from Wikidata[1] to effectively learn the posterior distribution of the prototype vectors of relations. ConceptFERE (Yang et al., 2021) boosts model performance by introducing the inherent concepts of entities[2]. The methods using external information face two obstacles. Firstly, it takes a lot of human effort to build the knowledge base. Secondly, the introduction of external knowledge may mislead the model to learn spurious correlations. The performance of the methods using external information is often not as good as that of the methods that only use relation descriptions. Hence, in this paper, we only deploy relation descriptions and do not introduce any external knowledge, and we aim at designing a more effective learning paradigm for FSRE.

## 3 Methodology

### 3.1 Problem formulation

**Definition 1** (*Relation Extraction (RE)*) Given a piece of text $d = (w_1, w_2, ..., w_n)$, a subject entity $\tilde{e}_s$ and an object entity $\tilde{e}_o$ are sequences of words in $d$, and the task of RE is to predict the relation $r \in \mathcal{R}$ between $\tilde{e}_s$ and $\tilde{e}_o$, where $R = \{r_1, ..., r_{|R|}\}$ is a predefined relation set.

Few-shot Relation Extraction (FSRE) typically follows an $N$-way-$K$-shot setting consisting of many individual $N$-way-$K$-shot relation extraction tasks. In each individual $N$-way-$K$-shot task, there are a support set $\mathcal{S}$ and a query set $\mathcal{Q}$. $\mathcal{S}$ contains $N$ novel relation classes, each class with $K$ labeled instances. The $N$-way-$K$-shot task aims to predict the implied relation $r \in \mathcal{R}_\mathcal{S}$ mentioned in an instance $q_i \in \mathcal{Q}$, where $\mathcal{R}_\mathcal{S} = \{r_1, ..., r_N\}$ denotes the relation set containing all relations mentioned in $\mathcal{S}$ and is different in different $N$-way-$K$-shot tasks.

It is a big challenge to directly utilize the support set to predict relations in query instances without any help. To address this issue, the researchers propose the meta-learning-based methods to utilize a large-scale auxiliary dataset $D_{base}$ which contains abundant labeled instances whose relations are disjoint from those in the testing tasks. The general idea behind this setting is to utilize the auxiliary dataset $D_{base}$ to help the model learn useful knowledge for FSRE.

The meta-learning-based methods commonly follow the $N$-way-$K$-shot setting on $D_{base}$ at the training stage. These methods continuously sample instances from $D_{base}$ and build $N$-way-$K$-shot tasks for training models. During training, the models can learn transferable knowledge that is valid for both current tasks and future tasks. The key to improving FSRE models is how to mine more general knowledge that can improve both current tasks and future tasks with relations unseen in $D_{base}$, rather than the knowledge specific to current tasks in $D_{base}$ only. Moreover, how to reach the optimal point in every single $N$-way-$K$-shot task is still an unexplored problem to be solved. Our proposed graph-based model generation (GM_GEN) framework also belongs to meta-learning-based methods and it is developed to address the above two issues.

### 3.2 Model overview

An overview of our proposed GM_GEN framework is shown in Fig. 2. It consists of 3 components: (1) An encoder based on the pre-trained language model (PLM). (2) A general graph-based generation module that generates different classification models based on different inputs (support, query samples, and relation descriptions) and the topology information which are considered as task descriptions. (3) Many task-specific models are generated for predicting the relations contained in the query samples. In our framework, different classification models generated by the same graph-based generation module can handle different $N$-way-$K$-shot tasks.

### 3.3 Encoder

Various types of encoders like convolutional neural network, recurrent neural network, and graph neural network, have been proposed for feature

[1] https://www.wikidata.org/

[2] https://concept.research.microsoft.com/Home/Download

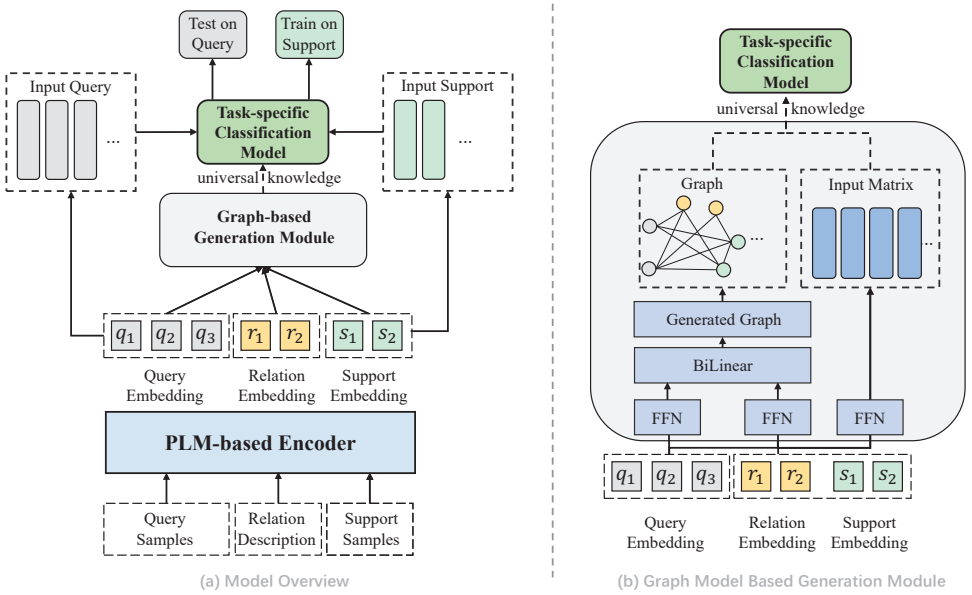



Figure 2: An overview of our proposed graph-based model generation (GM_GEN) framework.

extraction (Zeng et al., 2015; Soares et al., 2019; Peng et al., 2020). Among these, the pre-trained language model (PLM) based encoders achieve the best results since they contain a large amount of generalized semantic knowledge. Following existing studies (Soares et al., 2019; Han et al., 2021a; Liu et al., 2022), we employ BERT as the encoder to encode all the samples and relation descriptions. Specifically, for each sentence $d = (w_1, w_2, ..., w_n)$, we insert the '[CLS]' and '[SEP]' at the beginning and the end of the input sentence, respectively. In addition, we add four special tokens '[E1][/E1]' and '[E2][/E2]' at the beginning and end of the head entity and tail entity for the RE task, respectively. Such a setting is also consistent with those in (Soares et al., 2019; Han et al., 2021a; Liu et al., 2022; Han et al., 2021b).

We then concatenate the BERT-encoded vectors at the corresponding positions of '[E1]' and '[E2]' to obtain the sentence representation. It is formulated as follows:

$$\boldsymbol{x} = \boldsymbol{h}_i \oplus \boldsymbol{h}_j, \tag{1}$$

where $i$ and $j$ denotes the position of the special marker '[E1]' and '[E2]', and '$\oplus$' represents the concatenation operation. Since the relationship description text does not contain any entity, we concatenate the BERT-encoded embedding $\boldsymbol{h_0}$ on '[CLS]' and the mean of the embedding corresponding to the tokens after the '[CLS]' as the representation of the relation description. Specifically, the representation of a relation is defined as:

$$\boldsymbol{x_{rel}} = \boldsymbol{h}_0 \oplus \frac{\sum_{j=1}^{N} \boldsymbol{h}_i}{N}, \tag{2}$$

where $h_0$ represents the output of the '[cls]', and the following sum and division process denote the average of the token information other than the '[cls]' in relation description. $N$ is the number of tokens.

### 3.4 Graph-based generation module

The FSRE task contains many individual $N$-way-$K$-shot tasks. Existing methods follow the 'one-for-all' ideas, i.e., one general model for all tasks. However, such a general model trained on all tasks may contain a lot of knowledge irrelevant to the individual $N$-way-$K$-shot task. As a result, it is hard for existing 'one-for-all' methods to achieve the optimal point on each individual task. In view of this, we propose to employ model generation to decouple the complexity of the whole task space from that of each task space by solving each individual task under the 'one-for-one' paradigm, i.e., one task-specific model for one task in a more intuitive view. To this end, we design a graph-based generation module that can adapt well to the task space by combining the topology information and attribute features.

Specifically, we first design a bilinear transform layer to encode the topology information contained in an adjacency matrix $A$. Each node $i$ in $A$ denotes a support instance, a query sample, or a relation description. An edge $a_{ij}$ between two nodes $i$ and

$j$ in the adjacency matrix $A$ represents the semantic connection between them and can provide the clue for the classification of the current task. By absorbing the topology information, the generation module can better understand the current task. Formally, let $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ denote the embedding of the $i$-th and the $j$-th node, we calculate the edge score $a_{ij}$ using a bilinear transformation as follows:

$$a_{ij} = \begin{cases} \text{bilinear}(\text{FFN}(\boldsymbol{x}_i), \text{FFN}(\boldsymbol{x}_j)), & i \neq j \\ 0, & i = j \end{cases}, \tag{3}$$

where FFN means a feed-forward network.

The output of the graph convolution operation is as follows:

$$\boldsymbol{z}_i = \sigma(\sum_{j=1}^{n} \boldsymbol{A}_{ij} \boldsymbol{W} \boldsymbol{x}_j + b), \tag{4}$$

where $W$ and $b$ are the weight matrix and bias vector, respectively, and $\sigma$ denotes the ReLU function. Note the edges between support instances are different from those between the support instances and the query instance. The former provides information about relations, while the latter provides information about query samples. We hence propose to concatenate these two types of information and output these features as the weight matrix of the classification model. For each relation $r$ of the support set, we generate one output

$$\boldsymbol{o_r} = \boldsymbol{z}_r \oplus \frac{\sum_{j \in S_r} \boldsymbol{z}_j^s}{n}, \tag{5}$$

for each of the relations of the support set of the query, where $\boldsymbol{z}_r$ and $\boldsymbol{z}_j^s$ denote the embedding of the relation $r$ and the support instances, respectively, and $S_r$ denotes the support instances set with the relation $r$. The final output matrix $\boldsymbol{O} = \{o_1, ..., o_{|r|}\}$ contains general knowledge for inferring subsequent tasks and will be treated as the weight matrix to pass the universal knowledge from the general model to the task-specific classification models.

### 3.5 Generated classification model

The output matrix $\boldsymbol{O} = \{o_1, ..., o_{|r|}\}$ is generated by the model generation module and is now passed to each FSRE classification task. It contains universal knowledge which might be useful but is not specific enough for the classification task. Therefore, we further employ gradient updates so that the model can reach the optimal point for the current task. Specifically, we first input the query embedding $\boldsymbol{x}^q$ into the FFN and concatenate it with the original vector to obtain the classification feature vector $\boldsymbol{C^q}$, as defined as follows:

$$\boldsymbol{C^q} = \boldsymbol{x}^q \oplus \text{FFN}(\boldsymbol{x}^q). \tag{6}$$

The formula for the final classification is as follows:

$$\tilde{\boldsymbol{y}}_q = \sigma(\boldsymbol{O}\boldsymbol{C^q} + b). \tag{7}$$

During training on $D_{base}$, we use the cross-entropy loss computed for the query samples to calculate the gradient and process a gradient update on the encoder and the model generation module. During testing on each FSRE task, we only fine-tune the generated model for reaching the optimal point by calculating the cross-entropy on support instances. The loss in our framework is defined as follows:

$$\mathcal{L} = -\sum_{q \in D_{base}} \boldsymbol{y}_q \cdot \log(\tilde{\boldsymbol{y}}_q), \tag{8}$$

where $\boldsymbol{y}_q$ is the true label of query instance in $D_{base}$.

## 4 Experiment

### 4.1 Dataset

We use two commonly-used public datasets for our experiments: FewRel 1.0 and FewRel 2.0 (Han et al., 2018; Gao et al., 2019b). Both datasets are proposed for the FSRE task while FewRel 1.0 and FewRel 2.0 focus on in-domain (trained and tested on the same Wikipedia domain) and out-of-domain (trained on the Wikipedia domain and tested on a different biomedical domain) problems, respectively.

The FewRel 1.0 contains 70,000 sentences for 100 relations where each relation contains 700 instances. 100 relations are divided into 64, 16, and 20 splits to serve as the train, validation, and test set[3], respectively. Based on the training set of FewRel 1.0, FewRel 2.0 provides 10*100 validation samples and 10,000 test samples from a different domain. Also, note that FewRel 1.0 provides relationship description information while FewRel 2.0 does not. In general, FSRE tasks on FewRel 2.0 are more difficult than those on FewRel 1.0.

[3]The test set is not public, and the test results of the model can only be evaluated on: https://competitions.codalab.org/competitions/27980

| Encoder | Model | 5-1 | 5-5 | 10-1 | 10-5 |
|---|---|---|---|---|---|
| CNN | Proto-CNN (Snell et al., 2017b) | 72.65/74.52 | 86.15/88.40 | 60.13/62.38 | 76.20/80.45 |
| | Proto-HATT (Gao et al., 2019a) | 75.01/– – | 87.09/90.12 | 62.48/– – | 77.50/83.05 |
| | MLMAN(Ye and Ling, 2019) | 79.01/82.98 | 88.86/92.66 | 67.37/75.59 | 80.07/87.29 |
| BERT | Proto-BERT (Han et al., 2018) | 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 | 86.21/83.55 | 73.20/83.17 | 86.06/88.51 |
| | GNN (Satorras and Estrach, 2018) | – –/75.66 | – –/89.06 | – –/70.08 | – –/76.93 |
| | BERT-PAIR (Gao et al., 2019b) | 85.66/88.32 | 89.48/93.22 | 76.84/80.63 | 81.76/87.02 |
| | REGRAB (Qu et al., 2020) | 87.95/90.30 | 92.54/94.25 | 80.26/84.09 | 86.72/89.93 |
| | TD-Proto (Yang et al., 2020) | – –/84.76 | – –/92.38 | – –/74.32 | – –/85.92 |
| | 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 |
| | SimpleFSRE (Liu et al., 2022) | 91.29/94.42 | 94.05/96.37 | 86.09/90.73 | 89.68/93.47 |
| | GM_GEN | **92.65/94.89** | **95.62/96.96** | **86.81/91.23** | **91.27/94.30** |
| BERT w/ P | MTB (Soares et al., 2019) | – –/91.10 | – –/95.40 | – –/84.30 | – –/91.80 |
| | CP (Peng et al., 2020) | – –/95.10 | – –/97.10 | – –/91.20 | – –/94.70 |
| | LDUR (Han et al., 2021b) | 87.21/90.40 | 94.86/96.95 | 80.34/84.68 | 91.36/94.15 |
| | 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) | **96.97/97.03** | **98.32/98.34** | **93.97/94.99** | **96.58/96.91** |

Table 1: Comparison results in terms of accuracy (%) for FSRE methods on FewRel 1.0 validation / test set.

| Model | 5-1 | 5-5 | 10-1 | 10-5 |
|---|---|---|---|---|
| Proto-CNN | 35.09 | 49.37 | 22.98 | 35.22 |
| Proto-BERT | 40.12 | 51.5 | 26.45 | 36.93 |
| Proto-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** |

Table 2: Comparison results in terms of accuracy (%) for FSRE methods on FewRel 2.0 validation / test set.

### 4.2 Compared methods

We compared our model with the existing 15 baselines. Based on the type of the encoder, we classify these baselines into three categories: CNN-based, BERT-based, and BERT with post-training task[4] (BERT w/ P).

Among the first three baselines, **Proto-CNN** (Snell et al., 2017b), **Proto-HATT** (Gao et al., 2019a), and **MLMAN** (Ye and Ling, 2019), are all based on prototypical networks with CNN as the encoder. Proto-HATT and MLMAN further utilize the attention mechanism and a matching method to obtain a more accurate prototype, respectively.

Among the next nine BERT-based models, **Proto-BERT** (Han et al., 2018) is based on the basic prototypical network. **MAML** (Finn et al., 2017) is an optimization-based learning method. **GNN** (Satorras and Estrach, 2018) defines a neural network for few-shot learning tasks that is trained end-to-end. **BERT-PAIR** (Gao et al., 2019b) pairs each query instance with all the supporting instances and measures the similarity between instances. The next five baselines in this category employ knowledge from external knowledge bases to enhance FSRE models. Specifically, **REGRAB** (Qu et al., 2020) and **ConceptFERE** (Yang et al., 2021) utilize the global relation graph and the inherent concepts of entities, respectively. TD-Proto introduces text descriptions of entities and relations from Wikidata. **HCRP** (Han et al., 2021a) distinguishes hard tasks from easy ones by introducing the relation description. **SimpleFSRE** (Liu et al., 2022) directly adds the embedding of relation description to the prototype representation.

The last five baselines are based on different post-training tasks. **MTB** (Soares et al., 2019) designs a post-training task named matching the blanks. **CP** (Peng et al., 2020) proposes an entity-masked contrastive post-training framework. **LDUR** (Han et al., 2021b) develops a supervised contrastive post-training method to learn a more discriminative representation. **HCRP (CP)** and **SimpleFSRE (CP)** are based on the post-trained encoder provided by CP.

For a fair comparison with the existing BERT-based and BERT w/ P baselines, we provide the BERT-based and CP-based results for our proposed GM-GEN framework, respectively.

### 4.3 Settings

Following existing methods, we adopt the uncased model of BERT-base and CP as the sentence encoder in our experiments. The BERT-base model

[4]To distinguish from the pre-training tasks, we call those additional tasks on the BERT model as post-training tasks.

| Model | 5-1 | 5-5 | 10-1 | 10-5 |
|---|---|---|---|---|
| GM_GEN | 94.89 | 96.96 | 91.23 | 94.30 |
| ADD_Base | 94.46 (0.43↓) | 96.18 (0.78↓) | 88.91 (2.32↓) | 93.43 (0.87↓) |
| Add_GEN | 94.55 (0.34↓) | 96.49 (0.47↓) | 90.65 (0.58↓) | 94.06 (0.24↓) |
| GM_CLS | 94.76 (0.13↓) | 96.53 (0.43↓) | 91.06 (0.17↓) | 93.61 (0.69↓) |

Table 3: Ablation results in terms of accuracy (%) on FewRel 1.0 test set. ↓ denotes a drop of F1 score.

consists of a 12-layer transformer module, and CP has the same structure but is further post-trained by contrastive learning. During training, we set the train iteration number and validation iteration number to 30,000 and 1,000, respectively. The batch size is set to 4. The learning rate for the generation module is 1e-5, and that for the generated classification model is 5e-2 and 2e-2 in 1-shot and 5-shot, respectively. Following the official evaluation setting, we adopt 5-way-1-shot, 5-way-5-shot, 10-way-1-shot, and 10-way-5-shot to measure the model performance on the validation and testing set. Our platform is a 24 GB NVIDIA RTX 3090 GPU.

### 4.4 Main results

The comparison results on FewRel 1.0 and 2.0 are respectively shown in TABLE 1 and TABLE 2. We divide the results on FewRel 1.0 into three parts according to the type of encoders. On FewRel 2.0, we directly utilize the settings and results in HCRP (Han et al., 2021a). We have the following important findings from the results.

(1) Our proposed GM_Gen model performs the best among all methods using the same encoder. Compared to the metric-based and optimization-based methods, our proposed method better solves each $N$-way-$K$-shot classification task. All baselines are 'one-for-all' models and thus have limitations when they are generalized to new tasks. In contrast, our GM_Gen framework can reach the optimal point for these new tasks. The trend becomes more obvious on FewRel 2.0 for an out-of-domain test. This further proves the effectiveness of our model generation framework.

(2) The utilization of post-training tasks can bring general improvements to the model due to the advantage of post-training techniques designed for relation extraction tasks. The CP model is better than the MTB model which means that the CP post-training is more suitable for the FSRE than MTB. Our GM_GEN (CP) model outperforms all CP-based methods, showing that it consistently outperforms other methods under different settings.

(3) The improvement of our GM_GEN (CP) framework becomes less obvious. The pre-training task of the BERT model determines that it contains a large amount of semantic knowledge, which is not always necessary and may be noisy for relation extraction (RE). The post-training task of the CP model makes it more suitable for the RE by filtering noisy information. As a result, such CP-based models can improve the performance of all bert-based models. The reason why the performance margin between our proposed CP-based GM_GEN and other baseline methods becomes tighter is that the model is prone to overfitting under the setting of the CP-based FSRE task. The number of training epochs for CP-based models is usually much smaller than that of BERT-based models, i.e., the training of a CP-based model is not very sufficient.

(4) SimpleFSRE is superior to many baselines with complex designs, such as HCRP and LDUR. That is to say, the generalization ability of complex models is not always better than that of simple models. The complex structure of the models might be harmful to the performance and leads to overfitting.

(5) The models that utilize external knowledge, such as REGRAB and ConceptFERE, are not as good as expected. Extrinsic knowledge may bring noise besides the useful knowledge information, which affects the model in FSRE tasks. How to make better use of the knowledge information is still a challenging problem.

## 5 Analysis

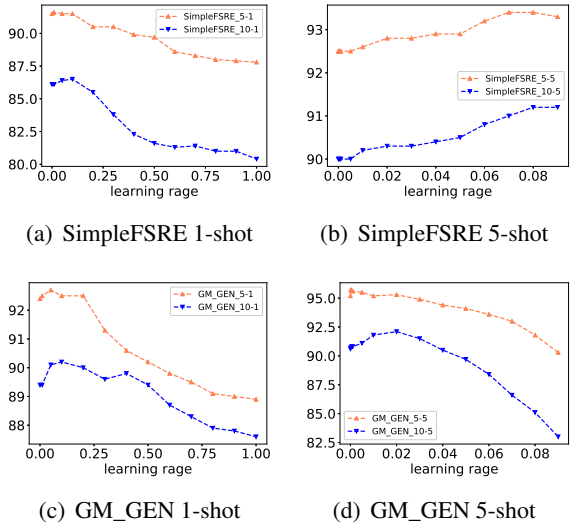
To get a deep insight into the proposed GM_GEN model, we conduct the ablation study, an investigation of hyper-parameter, a complexity analysis, and a visualization experiment.

### 5.1 Ablation study

We design three ablation experiments on FewRel 1.0 dataset to examine the influence of the graph-based generation module and the generated classification models. The results of the ablation study are summarized in Table 3. We detail the variants and analyze their effects as follows.

| | 5-1 | | 10-5 | |
|---|---|---|---|---|
| | time | space | time | space |
| GM_GEN | 5264.2S | 111.8M | 15533.4S | 111.8M |
| GNN | 5314.7S | 110.3M | 18335.9S | 110.3M |
| SimpleFSRE | 5256.1S | 109.8M | 15108.7S | 109.8M |
| HCRP | 5204.6S | 110.7M | 14573.3S | 110.7M |
| Proto-BERT | 4550.0S | 109.5M | 10612.2S | 109.5M |

Table 4: Complexity analysis. $S = Second$, $M = 1 \times 10^6$.




(a) SimpleFSRE 1-shot (b) SimpleFSRE 5-shot

(c) GM_GEN 1-shot (d) GM_GEN 5-shot

Figure 3: Impacts of the learning rate for the generated models on FewRel 1.0 validation set.

(1) "ADD_Base" only employs the addition operation on our framework without fine-tuning the classification model during the test phase. With this experiment, we show the performance of the simplest model in our framework.

(2) "Add_GEN" replaces the graph convolution process in the proposed GM_GEN with a direct addition operation on support instances and relation description. In this case, we wish to examine the influence of topology information on data. The results indicate that the deployment of topology information based on graph convolution operation is conducive to the generation module in our framework.

(3) "GM_CLS" employs the graph-based generation module but does not fine-tune the generated classification model. In this case, the classification model is not separated from the general generation module. Therefore, "GM_CLS" still follows the 'one-for-all' paradigm. Through this experiment, we wish to see the impact of decoupling the complexity of the whole task space from the complexity of each task space. The experiment results verify the necessity of decoupling operation.

## 5.2 Parameter analysis

The introduction of the model generation module brings an additional step of gradient update for our framework in the test phase of FSRE, which can affect the performance of our method. In this subsection, we provide an analysis of the learning rate of our generated models. The impacts of the learning rate on FewRel 1.0 utilized by Add_GEN and GM_GEN are shown in Figure 3. We have the following observations.

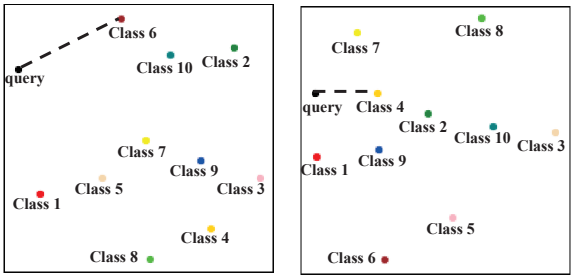
(1) The learning rate of the generated models can affect the classification performance, which also proves that our decoupling operation can improve the performance of the model when a suitable learning rate is adopted.

(2) The optimal learning rate is different under different situations. For example, in the case of 1-shot, the model performs the best at around a 0.05 learning rate, while in the case of 5-shot, it is between 0.6 and 0.9.

(3) The improvement of the model on the 5-shot setting is larger than that on the 1-shot setting because the gradient computed by the model is not general enough in the 1-shot setting where each class contains only one support sample. A too-large learning rate for the 1-shot setting will lead to overfitting on the single support sample, which further drops the performance of the model.

## 5.3 Analysis on computation cost

To demonstrate that the model improvement is not determined by the number of model parameters, we present the complexity analysis in Table 4.




(a) Before fine-tuning (b) After fine-tuning

Figure 4: Impacts of the fine-tuning on the generated models for FewRel 1.0 validation set. The dark dot represents the query sample, and other dots represent the prototypes of different classes. Note that the class 4 is the correct relation.

From the point of view of training parameters, Proto-BERT has the shortest running time and the fewest model trainable parameters because it only contains a pre-trained encoder and a simple dis-

tance metric. From another perspective in terms of running time, we find that GNN is also the longest-running time because it only processes one query sample per iteration.

Overall, the difference in the number of parameters and the running time for all methods is not very significant. However, the performance of our GM_GEN method is much better than these baselines.

### 5.4 Visualization on generated model

To more intuitively compare the effects of fine-tuning the classification model on different FSRE tasks, we utilize the t-SNE (van der Maaten and Hinton, 2008) to visualize the relationship between the query samples and the relation prototypes contained in the classifier before and after fine-tuning. From Figure 4, we can find that after the fine-tuning, the distance between the query and the prototypes of all classes becomes smaller, showing that the classifier after the adjustment is more suitable for the current classification task. Moreover, the model after fine-tuning can successfully pull the query point and the correct relation prototype closer.

## 6 Conclusions

To overcome the limitation of the popular 'one-for-all' learning scheme in current FSRE methods, we propose a novel graph-based model generation framework to decouple the complexity of the entire task space from that of each task space. By fine-tuning the generated models on the specific classification task, our method can naturally reach the optimal point in each task. Extensive experiments demonstrate the effectiveness of our proposed framework over existing baselines.

## 7 Limitations

While our model achieves a new state-of-the-art performance, it still has several limitations. Firstly, at the FSRE test phase, the learning rate used to fine-tune the generated model in our proposed framework needs to be defined in advance using the validation set. Secondly, the classification model generated in our framework is simple. Though effective, we believe more advanced classification models can further improve the performance, which can be a future direction.

## Acknowledgments

We thank the anonymous reviewers for their valuable comments. The work described in this paper is supported by the NSFC project (62276193).

## References

Antreas Antoniou, Harrison Edwards, and Amos J. Storkey. 2019. How to train your MAML. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

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, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019a. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6407–6414.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019b. Fewrel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6249–6254.

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2605–2616.

Jiale Han, Bo Cheng, and Guoshun Nan. 2021b. Learning discriminative and unbiased representations for few-shot relation extraction. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 638–648.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 4803–4809.

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

Huai-Yu Li, Weiming Dong, Xing Mei, Chongyang Ma, Feiyue Huang, and Bao-Gang Hu. 2019. Lgm-net: Learning to generate matching networks for few-shot learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3825–3834.

Wanli Li, Tieyun Qian, Ming Zhong, and Xu Chen. 2022. Interactive lexical and semantic graphs for semisupervised relation extraction. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12.

Hongtao Lin, Jun Yan, Meng Qu, and Xiang Ren. 2019. Learning dual retrieval module for semi-supervised relation extraction. In *The World Wide Web Conference, WWW*, pages 1073–1083.

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, Dublin, Ireland, May 22-27, 2022*, pages 757–763. Association for Computational Linguistics.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *CoRR*, abs/1803.02999.

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 2020, Online, November 16-20, 2020*, pages 3661–3672.

Meng Qu, Tianyu Gao, Louis-Pascal A. C. Xhonneux, and Jian Tang. 2020. Few-shot relation extraction via bayesian meta-learning on relation graphs. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7867–7876.

Melanie Reiplinger, Michael Wiegand, and Dietrich Klakow. 2014. Relation extraction for the food domain without labeled training data - is distant supervision the best solution? In *Advances in Natural Language Processing - 9th International Conference on NLP, PolTAL 2014*, volume 8686 of *Lecture Notes in Computer Science*, pages 345–357. Springer.

Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. 2019. Meta-learning with latent embedding optimization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Victor Garcia Satorras and Joan Bruna Estrach. 2018. Few-shot learning with graph neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017a. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017b. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087.

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. In *Journal of Machine Learning Research*, volume 9, pages 2579–2605.

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 *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 2273–2276.

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, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 987–991.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-level matching and aggregation network for few-shot relation classification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2872–2881.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP*, pages 1753–1762.

Andrey Zhmoginov, Mark Sandler, and Max Vladymyrov. 2022. Hypertransformer: Model generation for supervised and semi-supervised few-shot learning. *CoRR*, abs/2201.04182.