# Few-Shot Text Classification with Edge-Labeling Graph Neural Network-Based Prototypical Network

**Chen Lyu**
Peking University
chenlyu@pku.edu.cn

**Weijie Liu**
Peking University
dataliu@pku.edu.cn

**Ping Wang**
Peking University
pwang@pku.edu.cn

## Abstract

In this paper, we propose a new few-shot text classification method. Compared with supervised learning methods which require a large corpus of labeled documents, our method aims to make it possible to classify unlabeled text with few labeled data. To achieve this goal, we take advantage of advanced pre-trained language model to extract the semantic features of each document. Furthermore, we utilize an edge-labeling graph neural network to implicitly models the intra-cluster similarity and the inter-cluster dissimilarity of the documents. Finally, we take the results of the graph neural network as the input of a prototypical network to classify the unlabeled texts. We verify the effectiveness of our method on a sentiment analysis dataset and a relation classification dataset and achieve the state-of-the-art performance on both tasks.

## 1 Introduction

In recent years, deep learning has achieved great success in many fields such as computer vision, speech recognition and natural language processing. However, to train a deep learning model usually requires a large scale of manually annotated data, which greatly limits the practicality and scalability of the model.

In the text classification task, the problem of greed for large amounts of labeled data also exists. Text classification is an important task in natural language processing field. Over the years, many text classification methods have been proposed, varying from CNN based methods (Kim, 2014; Johnson and Zhang, 2015; Zhang et al., 2015) to RNN, including LSTM and GRU-based methods (Zhou et al., 2015; Yang et al., 2016). But all the above supervised learning methods require a large corpus of labeled data, making these models hindered in practical application.

In recent years, some scholars have contributed to solving the few-shot text classification problem. Han (2018) proposed a dataset FewRel and tested the performance of several typical few-shot learning models proposed in recent years on this dataset. Yu (2018) proposed an adaptive metric learning model, which can automatically determine the best weighted combination of a set of metrics obtained from a meta-learning process for a newly arrived few-shot text classification task. Gao (2019) proposed a prototypical network model (Snell et al., 2017) based on a hybrid attention mechanism, which has a better performance for noisy data scenes. Sun (2019) improved prototypical network by adopting hierarchical attention mechanism, which is applied in feature level, word level and instance level to enhance the expressive ability of semantic space. Geng (2019) applied the dynamic routing algorithm in meta-learning and proposed an induction network, which achieves a better generalization ability on different few-shot text classification tasks.

We believe that previous few-shot text classification methods generally have two flaws. First, all the methods mentioned above adopt a GloVe word embedding combined with CNN or RNN structure for text embedding instead of more advanced pre-trained language models proposed in recent years, such as ELMo (Peters et al., 2018), GPT (Radford et al., 2018) and BERT (Devlin et al., 2018). Second, these methods focus more on the semantic features of the texts itself, ignoring the potential relationships

*Proceedings of the 28th International Conference on Computational Linguistics*, pages 5547–5552
Barcelona, Spain (Online), December 8-13, 2020

between texts. In this paper we propose Edge-Labeling Graph Neural Network-Based Prototypical Network (EGNN-Proto) to further tackle the few-shot text classification task. EGNN-Proto takes advantage of the advanced pre-trained language model BERT, and improves the prototypical network (Snell et al., 2017) by combining an edge-labeling graph neural network (Kim et al., 2019) to better characterize and utilize the relationship between texts to achieve a better performance on few-shot text classification task. The main contributions of our work are as follows:

- We propose Edge-Labeling Graph Neural Network-Based Prototypical Network (EGNN-Proto) for few-shot text classification. To our knowledge, our model is the first to combine a graph neural network with a prototypical network and the first to utilize an edge-labeling graph neural network to solve the few-shot text classification problem.
- Our method outperforms the current state-of-the-art models on two few-shot text classification datasets, i.e. Amazon Review Sentiment Classification (ARSC) and Few-shot Relation classification dataset (FewRel).

## 2 Related Works

Few-shot learning (FSL) aims to learn classifiers for new classes with only a few training examples per class. Given a support set of labeled instances, the goal is to classify instances from query set. The seminal work on few-shot learning dates back to the early 2000s (Fei-Fei et al., 2003; Fei-Fei et al., 2006). More recent work can be divided into two types: similarity-based methods (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018) and optimization-based methods (Ravi and Larochelle, 2016; Munkhdalai and Yu, 2017). Researchers have also studied FSL in various NLP tasks (Yu et al., 2018; Gu et al., 2018; Gao et al., 2019; Sun et al., 2019; Geng et al., 2019; Hu et al., 2019).

Graph neural networks (GNN) were first proposed by Gori (2005) and Scarselli (2008). Li (2015) further extended it with gated recurrent units. Generalized convolution-based propagation rules also have been directly applied to graphs (Bruna et al., 2013; Henaff et al., 2015; Defferrard et al., 2016), offering a balance between expressivity and sample complexity. In recent years, A few approaches (Garcia and Bruna, 2017; Kim et al., 2019; Gidaris and Komodakis, 2019) further explored GNNs for few-shot learning.

## 3 Method

### 3.1 Few-shot Text Classification

In few-shot text classification task, we define the set of labeled samples as support set $S$ and the set of unlabeled samples as query set $Q$. Following previous works (Sun et al., 2019; Geng et al., 2019), we adopt the $N$-way, $k$-shot setting, where the support set $S$ contains $k$ labeled samples for each of $N$ classes, noting that $k$ is rather small in a few-shot learning task. Our goal is to perform meta-learning on the training set, then extract transferable knowledge that will allow us to deliver better few-shot learning on the support set $S$ and classify the samples from query set $Q$ as accurately as possible.

### 3.2 Method Overview

As illustrated in Figure 1, our model consists of three major components, i.e., text embedding component, edge-labeling graph neural network component and prototypical network component. Text embedding component is mainly used to extract semantic features from the text and transform the original text sequences into text embedding. Then we adopt edge-labeling graph neural network to model the intra-cluster similarities and inter-cluster dissimilarities of the texts, thus we can measure the underlying relationships between a text sample and another. Finally, we classify the samples from query set using a prototypical network.

### 3.3 Text Embedding

In a few-shot text classification task, only a small amount of annotated data can be used to train the classifier. So we choose to make use of a pre-trained language model to help use better extract the

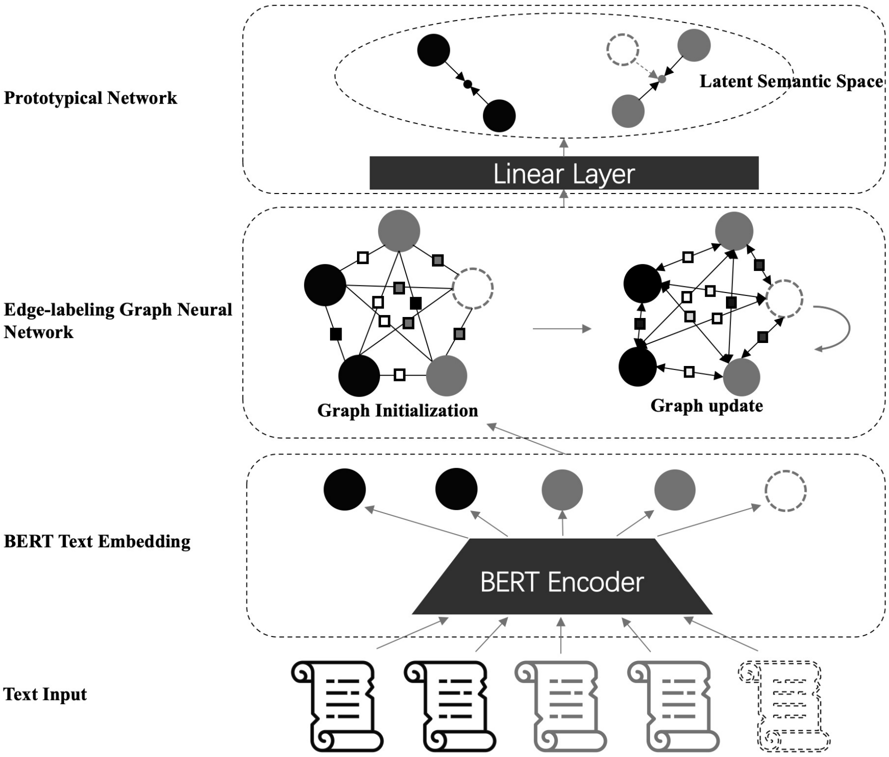

Figure 1: An Overview of EGNN-Proto. The example shows the workflow of a 2-way 2-shot few shot text classification task. Document icons in different colors represent text from different classes, and the icon formed by dotted lines represents the query text. Circles represent the text embeddings, and their shapes and colors have the same meaning as icons. The strength of edge feature is represented by the color in the square, following (Kim et al., 2019).

feature of the text. Here we adopt a transformer encoder from BERT (Devlin et al., 2018). A transformer encoder stack consists of 12 layers of encoders, each of which consists of a bidirectional self-attention layer and a fully connected layer. Every token in the input of the stack is first embedded into a learned $d$-dimensional embedding, and then transformed progressively every time it traverses one of the BERT Encoder layers. Theoretically, each word embedding output from each encoder layer contains the feature of the entire text. Here we adopt the word embedding of the special symbol $[CLS]$, which marks the beginning of a text, from the penultimate layer as the text embedding of the whole text $t$, represented as $e = f_{emb}(t|\theta_{emb})$, where $\theta_{emb}$ represents parameters in text embedding component, and are fine-tuned while training.

### 3.4 Edge-labeling Graph Neural Network

We introduce the edge-labeling graph neural network, which is initially proposed by Kim (2019) for few-shot image classification task, to better characterize the potential relationships between texts. Given the text embedding of all samples of a task, a fully connected graph is initially constructed, where each node represents each sample and each edge represents relationship between the connected nodes. Node features, which are initialized by the text embeddings of texts, i.e. $e$, and edge features, which are initialized based on whether the connected nodes belong to the same class, are alternately updated till convergence. Finally, the updated node features $e_f = f_{egnn}(e|\theta_{egnn})$ are obtained for further processing.

### 3.5 Prototypical Network

Instead of computing the prediction probability of query nodes directly through edge features and node features as Kim (2019), we introduce prototypical network (Snell et al., 2017) to classify samples from a more general perspective. Each sample feature $e_f$ is initially transformed through a linear layer to get a new feature representation $e_l = f_{linear}(e_f|\theta_{linear})$. Then prototypical network computes a prototype vector as the representation of each class, which is the mean vector representations of the support samples from the class. We compare the distances between all prototype vectors and a query vector, then classify the query sample to the nearest one.

## 4 Experiments

### 4.1 Implementation Details

In implementation, we adopt BERT-base as text encoder, thus a sample is represented by a 768-dimensional vector after text embedding. The episode training strategy that Vinyals (2016) proposed is adopted in training procedure.

### 4.2 Datasets

We evaluate our method and compare it with other methods on two few-shot task classification datasets, i.e. Amazon Review Sentiment Classification (ARSC) and Few-shot Relation classification dataset (FewRel).

ARSC dataset (Blitzer et al., 2007) consists of English reviews of 23 types of products on Amazon. Following the experiment setting of Yu (2018), we creat a 12-way 5-shot text classification task on this dataset.

FewRel (Han et al., 2018) is a large-scale supervised dataset consisting of 70000 instances on 100 relations derived from Wikipedia. In our experiment, we follow the settings of Sun (2019) and evaluate our method on 5-way 5-shot and 10-way 5-shot settings.

### 4.3 Results and Analysis

| Model | Mean Acc |
|---|---|
| Matching Networks$^\diamond$ | 65.73 |
| Prototypical Networks$^\diamond$ | 68.17 |
| Graph Network$^\diamond$ | 82.61 |
| Relation Networks$^\diamond$ | 83.07 |
| SNAIL$^\diamond$ | 82.57 |
| ROBUSTTC-FSL$^\diamond$ | 83.12 |
| Induction Networks$^\diamond$ | 85.62 |
| EGNN-Proto (ours) | **88.26** |

| Model | 5 way 5 shot | 10 way 5 shot |
|---|---|---|
| Finetune$^*$ | 68.66±0.41 | 55.04±0.31 |
| kNN$^*$ | 68.77±0.41 | 55.87±0.31 |
| MetaN$^*$ | 80.57±0.48 | 69.23±0.52 |
| GNN$^*$ | 81.28±0.62 | 64.02±0.77 |
| SNAIL$^*$ | 79.40±0.22 | 68.33±0.25 |
| Proto$^*$ | 89.05±0.09 | 81.46±0.13 |
| PHATT$^*$ | 90.12±0.04 | 83.05±0.05 |
| HAPN$^*$ | 91.02±0.11 | 84.16±0.19 |
| EGNN-Proto (ours) | **92.29±0.23** | **86.09±0.31** |

Table 1: Classification Accuracy (%) on ARSC (on the left) and FewRel Dataset (on the right). The results of methods $^\diamond$ are reported by Geng (2019) while $^*$ reported by Sun (2019).

Experiment results on ARSC and FewRel are presented in Table 1. The proposed EGNN-Proto model achieves an 88.26 % accuracy on ARSC dataset, outperforming the existing state-of-the-art model, Induction Networks, by a notable 2.6 % improvement. And EGNN-Proto also shows a great improvement of accuracy on FewRel dataset.

## 5 Conclusion

This work addresses the problem of few-shot text classification. We take advantage of advanced pre-trained language model BERT to extract the semantic feature of the text. Then we introduce edge-labeling graph neural network to further model the potential relationships between texts. Finally we utilize a prototypical network to classify the query text. In the experiments, our method achieves the state-of-the-art performance on both ARSC and FewRel datasets. For future work, we can consider generalizing our method to other few-shot NLP problems, and even to non-NLP tasks.

## References

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2013. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2003. A bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings of the Ninth IEEE International Conference on Computer Vision-Volume 2*, page 1134.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6407–6414.

Victor Garcia and Joan Bruna. 2017. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction networks for few-shot text 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 3895–3904.

Spyros Gidaris and Nikos Komodakis. 2019. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–30.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. 2005. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734. IEEE.

Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631.

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.

Mikael Henaff, Joan Bruna, and Yann LeCun. 2015. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*.

Ziniu Hu, Ting Chen, Kai-Wei Chang, and Yizhou Sun. 2019. Few-shot representation learning for out-of-vocabulary words. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Rie Johnson and Tong Zhang. 2015. Effective use of word order for text categorization with convolutional neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–112.

Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. 2019. Edge-labeling graph neural network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.

Tsendsuren Munkhdalai and Hong Yu. 2017. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2554–2563.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087.

Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. 2019. Hierarchical attention prototypical networks for few-shot text 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 476–485.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1199–1208. IEEE.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse few-shot text classification with multiple metrics. In *NAACL-HLT*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. 2015. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*.