Neighbour Contrastive Learning with Heterogeneous Graph Attention Networks on Short Text Classification

Anonymous ACL submission

Abstract

Graph neural networks (GNNs) have attracted extensive research interests in text classification tasks, due to their superiority in representation learning. However, most existing studies adopt the same semi-supervised learning setting as the vanilla Graph Convolution Network (GCN), which require a large amount of labelled data during training and thus are less robust when dealing with large-scale graph data with few labels. Additionally, graph structure information is normally captured by direct information aggregation via network schema and missing adjacency knowledge may hinder the performance. Addressing those problems, this paper proposes a novel method to learn graph structure, by using simple neighbour contrastive learning for an existing self-supervised heterogeneous graph neural network model (NC-HGAT). It considers the graph structure information from heterogeneous graphs with a multi-layer perceptrons (MLPs) and delivers consistent results, despite the corrupted neighbouring connections. Extensive experiments have been implemented on four benchmark short-text datasets, and demonstrate that our proposed model NC-HGAT outperforms the state-of-the-art methods on three datasets and achieves a competitive result on the remaining dataset.

1 Introduction

Text classification, is a fundamental task in natural language processing (NLP), which can be applied into a variety of downstream tasks, such as question answering, machine translation and sentiment analysis (Li et al., 2020). The representation learning ability of textual features is a leading cause for the performance of models on text classification, and consequently, it is a pressing need to study how to extract textual features more effective. Recently, graph neural networks (GNNs) have been increasingly applied to text classification tasks due to their advantage of dealing with complex semantics and topological information, by modelling texts as graph structure (Wu et al., 2020). Different from most long text classification studies, we mainly focus on short text classification, as our daily communication is increasingly completed via short texts, such as tweets, messenger and online comments, and thus it is important to study this field, specifically.

Most existing studies of GNNs on text classification tasks are trained in a semi-supervised manner, same as the vanilla Graph Convolution Network (GCN) (Kipf and Welling, 2016), which requires sufficient labelled data and cannot be satisfied in many real scenarios. Therefore, the shortage of labelled data may undermine the performances of graph neural network models on classification tasks, particularly with large scale data (Linmei et al., 2019; Sun et al., 2021).

On the other hand, most graph-based learning models only capture one-hop neighbourhoods and the associated textual features by supervised information aggregation, which may not be able to incorporate the high-order, rich relations among texts (Liu et al., 2021), and is not robust when the connections among nodes are noisy or missing (Hu et al., 2021).

To address the above problems, we propose to integrate neighbouring contrastive learning with the heterogeneous graph attention network (NC-HGAT). Contrastive learning can learn intrinsic and transferable topological information, enhance the performance of graph neural networks (Qiu et al., 2020) and is widely applied in NLP tasks for pre-training (Gunel et al., 2020). The neighbouring contrast learning enables the proposed model to transform k^{th} structural-aware features, without direct message-passing modules and hence improve the robustness despite the missing connections between words during inference (Hu et al., 2021), with limited labelled data.

The contributions of the paper are summarised

as follows:

- To the best of our knowledge, *this is the first attempt* to apply contrastive learning with a heterogeneous graph neural network on short text classification tasks.
- We propose to use a simple MLP to learn the neighbouring information, without direct message-passing, which can be easily applied to existing graph neural network models (Hu et al., 2021) on text classification.
- Experimental results on three of four datasets prove the outperformance of the proposed model on short text classification over the state-of-the-art with limited labelled data, and it also delivers a competitive result on the remaining dataset.

2 Related Work

Extensive studies have been conducted on text classification, such as traditional machine learning using manually designed features (Blei et al., 2003), convolutional neural networks (Chen, 2015) and recurrent neural networks (Liu et al., 2015). Recently, graph neural networks (GNNs) showed promising performance on text classification, as text can be modelled as edges and nodes in a graph structure. For instance, TextGCN (Wang et al., 2019) applied the vanilla GCN to heterogeneous graphs, on graphs built from a text corpus, and gained improved results. (Linmei et al., 2019) proposed a novel heterogeneous graph attention networks model (HGAT) with a dual attention mechanism, to consider more relations of different nodes. Recently, (Yang et al., 2021) introduced an orphan category to HGAT, to remove unrelated stop-words, and improve classification accuracy. (Liu et al., 2021) also incorporated the attention mechanism with deep diffusion layers, to enrich the context information of texts. However, these methods all relied heavily on the direct message-passing function to learn node feature transformation, and the performance will be undermined when labelled training data is limited. We propose, for the first time to the best of our knowledge, to solve the problem by applying contrastive learning of graph structure in the text classification tasks.

3 Model

In this section, we will introduce our NC-HGAT model, which is mainly based on the HGAT model (Linmei et al., 2019) and the neighbouring contrastive learning adopted in the Graph-MLP model (Hu et al., 2021).

3.1 HGAT

Compared with TextGCN (Wang et al., 2019), which directly applies GCN to different subgraphs, HGAT introduce a dual attention mechanism: type-level attention and node-level attention, to learn the relative influence of the different types and neighbouring nodes on the target node during information aggregation (Linmei et al., 2019). The type-level attention a_t is calculated as:

$$a_t = softmax(\sigma(\mu_t \cdot [h_i||h_t])) \quad (1)$$

where σ is a LeakyReLU activation, μ_t denotes the attention of the type t of the node, and operation $||$ is a concatenation. h_i and h_t are the node and type embedding. Then a softmax function is applied to normalise all types of neighbours of node i. The node level attention $a_n$is formulated based on the type level attention a_t from Equation 1:

$$a_n = softmax(\sigma(v_t \cdot [h_i||h_{j'}])) \quad (2)$$

where v denotes the attention vector and $h_{j'}$ is the embedding of node j, which have already considered the type-level attention. The two attention mechanisms will be integrated into the heterogeneous graph convolution, to update the embedding of nodes in the next layer:

$$H^{l+1} = \sigma(\sum \hat{A}_t \cdot H_t^l \cdot W_t^l) \quad (3)$$

where $\hat{A}$ is an adjacency matrix with type t edges, H_t^l is the feature of type t neighbouring nodes of the target node and W_t^l is a weight matrix.

3.2 Neighbouring Contrastive Learning

The neighbouring contrastive learning is mainly implemented by calculating the contrastive loss for node i. The initiative behind it is that neighbouring documents are more likely to have a same class label. The node feature X will simply pass two linear layers with activation σ and layer normalisation LN, dropout in between to avoid over-fitting, given by (Hu et al., 2021):

$$Z = W^1[Dropout(LN(\sigma(XW^0)))] \quad (4)$$

Where W^1 and W^0 are the weight matrices of two layers. The number of linear layers could be set

differently (from 1-7) as analysed in 4.4. Next, the embedding Z will be used to calculate the neighbouring contrastive loss:

$$loss_{NC} = -log\frac{\sum_j \lambda exp(sim(z_i, z_j)/\eta)}{\sum_k exp(sim(z_i, z_k)/\eta)} \quad (5)$$

where λ is a connection measure of node j and i and is not zero only when the node j is within the k-hop neighbourhood of node i. sim is the cosine similarity and η is the temperature parameter.

3.3 Model Training

Considering limited labelled data is provided, we only use 20 labelled documents per class as training data. We firstly use the HGAT model to build graphs from the text corpus and learn the representation of nodes with the dual-level attention mechanism. At the same time, we use the MLP-based model to learn more graph structure information, without an explicit message-passing function. To be more specific, the k-hop neighbours are considered more similar to the target node and this k^{th} power of the neighbouring information is in the range of [1,2,3,4,5,6,7]. If the neighbouring node is not a k-hop of the target node, the neighbours' information would be considered zero. Then, we calculate the neighbouring contrastive loss, $loss_{NC}$.

$$loss_{total} = loss_{NLL} + \beta * loss_{NC} \quad (6)$$

The total loss of our model would be the sum of the conventional negative log-likelihood loss $loss_{NLL}$ and the contrastive loss, $loss_{NC}$. β is a coefficient parameter to balance the total loss. The gradient descent algorithm is applied to optimise the total loss.

4 Experiments

4.1 Dataset

We use the same four benchmark short text datasets as (Linmei et al., 2019), and the details are as follows. The movie review dataset (MR) (Pang and Lee, 2005) has 5,331 positive and 5,331 negative reviews, where each review is one sentence. Twitter, a sentiment classification dataset provided by the NLTK library of Python, contains 5,000 positive and negative tweets, respectively. Ohsumed, is provided by (Yao et al., 2019) where a graph convolution network model is applied for text classification. AGNews are randomly selected 6,000 news from (Zhang et al., 2015). We do not have results on the other two datasets, Snippets and Tagmynews, used by (Linmei et al., 2019; Yang et al., 2021), due to the memory limit of the GPU.

4.2 Baselines and Experiment Settings

Baselines We consider three widely applied NLP models and other three graph neural network models, applied as baselines for text classification.

SVM +TFIDF and SVM + LDA are conventional machine learning classifiers, using classic features, including TF-IDF and LDA features (Salton and Buckley, 1988; Blei et al., 2003).

BERT, deploying a bidirectional Transformer encoder (Devlin et al., 2018), is a widely-applied model in NLP.

TextGCN is the first study which applies GCN to text, by building heterogeneous graphs from a text corpus (Yao et al., 2019).

HAN considers the importance of both node and meta-path, by introducing an attention mechanism into the heterogeneous graph neural network (Wang et al., 2019).

HGAT integrating a dual attention mechanism into heterogeneous information network (Linmei et al., 2019; Yang et al., 2021), is state of the art on the short text classification tasks.

Experiment Settings The hyper-parameters of NC-HGAT are mainly borrowed from the experiments of HGAT (Linmei et al., 2019) and GraphMLP (Hu et al., 2021). 40 labelled documents per class are randomly selected and split equally into training and validation sets. We use two layers and the number of hidden units is 512, learning rate 0.005, with an 80% dropout rate at each layer. The dimension of pre-trained word embeddings is set to 100. The k^{th} power of adjacency matrix, temperature parameter η and the coefficient balance parameter β are set by using grid search. The range of η and β are [0,1,2] and [0.5, 1.0, 2.0, 3.0], respectively.

4.3 Experimental Results

Figure 1 shows the classification performance of different models on the four benchmark datasets. The proposed model NC-HGAT outperforms all baselines on three datasets, demonstrating the effectiveness of the neighbouring contrastive learning with the heterogeneous graph attention network on short text classification. The minor underperformance of NC-HGAT on the MR dataset may

be because it captures more background information or stop-words, which are unrelated to a specific class, thus diminishing the result.

Dataset	Evaluation	AGNews	MR	Ohusmed	Twitter
SVM+TFIDF	Accuracy	59.45	54.29	39.02	53.69
	F1-score	59.79	48.13	24.78	52.45
SVM+LDA	Accuracy	65.16	54.40	38.61	54.34
	F1-score	64.79	48.39	25.03	53.97
Bert	Accuracy	69.45	53.48	21.76	52.00
	F1-score	69.31	46.99	4.81	43.34
Text-GCN	Accuracy	67.61	59.12	41.56	60.15
	F1-score	67.12	58.98	27.43	59.82
HAN	Acurracy	62.64	57.11	36.97	53.75
	F1-score	61.23	56.46	26.88	53.09
HGAT	Acurracy	72.10	**62.75**	42.68	63.21
	F1-score	71.61	**62.36**	24.82	62.48
NC-HGAT	Accuracy	**73.15**	62.46	**43.27**	**63.76**
	F1-score	**72.06**	62.14	**27.98**	**62.94**

Figure 1: Models Evaluation on Four Datasets

Number of Layers	Accuracy(%)	F1-score (%)
1	63.04	62.99
2	61.86	61.22
3	61.05	60.93
4	63.66	62.5
5	63.28	62.63
6	**63.76**	**62.9**
7	62.79	62.28

Figure 2: Model Performance with Different Layers on the Twitter Dataset

4.4 Impact of Layer Numbers of MLP

To investigate the impact of the MLP layer number deployed in section 3.2, we evaluate our NC-HGAT model with 1-7 layers on the Twitter and AGnews datasets. As shown in Figures 2, 3, the model with two layers performs better on the AGnews dataset; for the Twitter dataset, six layers perform the best. As for the news dataset, if the number of layers is excessive, the vanishing gradient and over-processed information will lead to an unstable model. The node representations may also become indistinguishable, known as the over-smoothing problem (Yang et al., 2020). While for the Twitter dataset, distant words may still be able to classify the document and six layers can capture sufficient structural information.

Number of Layers	Accuracy(%)	F1-score(%)
1	73.00	71.72
2	**73.15**	**72.06**
3	72.50	71.81
4	72.85	71.61
5	72.50	71.03
6	72.60	71.16
7	72.3	71.45

Figure 3: Model Performance with Different Layers on the AGnews Dataset

5 Conclusion and Future Work

In this paper, we propose to use contrastive learning to capture the topological information with HGAT on short text classification tasks. Extensive experiments illustrate that neighbour contrastive learning effectively learns and integrates structural information among entities and thus enhances the robustness of the existing model, particularly when there are limited labelled data. There may exist some better contrastive learning for graph structure methods, which we will explore in future work.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

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*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Yang Hu, Haoxuan You, Zhecan Wang, Zhicheng Wang, Erjin Zhou, and Yue Gao. 2021. Graph-mlp: Node classification without message passing in graph. *arXiv preprint arXiv:2106.04051*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2020. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*.

Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short 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 4821–4830.

Pengfei Liu, Xipeng Qiu, Xinchi Chen, Shiyu Wu, and Xuan-Jing Huang. 2015. Multi-timescale long short-term memory neural network for modelling sentences and documents. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2326–2335.

Yonghao Liu, Renchu Guan, Fausto Giunchiglia, Yanchun Liang, and Xiaoyue Feng. 2021. Deep attention diffusion graph neural networks for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8142–8152.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*.

Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. 2020. Gcc: Graph contrastive coding for graph neural network pre-training. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1150–1160.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Zhongtian Sun, Anoushka Harit, Jialin Yu, Alexandra I Cristea, and Noura Al Moubayed. 2021. A generative bayesian graph attention network for semi-supervised classification on scarce data. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *The World Wide Web Conference*, pages 2022–2032.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.

Chaoqi Yang, Ruijie Wang, Shuochao Yao, Shengzhong Liu, and Tarek Abdelzaher. 2020. Revisiting over-smoothing in deep gcns. *arXiv preprint arXiv:2003.13663*.

Tianchi Yang, Linmei Hu, Chuan Shi, Houye Ji, Xiaoli Li, and Liqiang Nie. 2021. Hgat: Heterogeneous graph attention networks for semi-supervised short text classification. *ACM Transactions on Information Systems (TOIS)*, 39(3):1–29.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.