Chinese short text classification by combining Bert and graph convolutional network

Mo Chen[a], Chunlong Yao*[a], Xu Li[b], Lan Shen[a]
[a]School of Information Science and Engineering, Dalian Polytechnic University, Dalian, Liaoning, China; [b]Engineering Training Center, Dalian Polytechnic University, Dalian, Liaoning, China

ABSTRACT

Short Text Classification is the fundamental task in the nature language processing. There is a lack of language structure and uneven classification of data samples in short texts, which limit the development of deep learning based short text classification. To address the limitations of text sequences, we propose using a large-scale pre-trained language model Bert to obtain feature information between words and bureaus in the text, Graph Convolutional Network (GCN) with double-layer convolutional network can obtain the dependency relationships between words. We propose to combine Bert with GCN in short Chinese medical texts, where BertGCN outperforms better than other's methods in classification accuracy.

Keywords: Short text classification, Bert, GCN

1. INTRODUCTION

Text classification is an important and classical problem in natural language processing (NLP), which is widely used in e-commerce[1], social media[2] and healthcare. Text classification of Chinese medical data consists of long text data and short text data, with the long text containing clinical data and personal consultation records. Most of the short text data are user questions collected from medical Q&A systems, and the questions are classified to obtain the user's question intent. The use of deep learning methods to obtain document contextual information and key node information can help improve the accuracy of text classification.

Compared with other's classical text classification, short text classification poses additional challenges. Firstly, short texts in Chinese medical typically involve sentences with an average length from 15 to 30 words, reducing the information that can be obtained from text-based features. Secondly，Unlike topic classification, there are too many labels and less of data in Chinese medical, in which improves keyword extraction in order to increasing the accuracy of the text classification. For example: The label "Treatment options" can be classified as "Exercise", "Diet" and "Bedtime". All these factors make it difficult to separate user's intention by purely relying on text which is heterogeneous and contains less data.

In this work, we propose a model BertGCN which enhance the textual information by combining the Bert which is large-scale pretraining and GCN which is transudative learning. A large-scale heterogeneous graph is constructed as a corpus for BertGCN, where the nodes are words. The nodes are pre-trained with Bert to obtain an initialised feature representation, and use GCN to capture structural and semantic features of text. The two predictions fused and used as the basis for text classification. Accurately, compared with other studies our study improved accuracy by 1.21%, 10.39% respectively on the dataset.

2. RELATED WORKS

In text classification, the classical models are based on CNN, RNN, Transformer and GCN. Based CNN model can get the text local information. Based RNN model adds sequence information to make text read in sequence. Based Transformer model uses multi-head self attention to get more text global information. Based GCN model use graph to show the information and the local information. In order to obtain contextual information and local feature information

* Yaocl@dlpu.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125060O
 · 0277-786X · doi: 10.1117/12.2662199

from the text and to reduce the problem of short and separately uneven sample sequences in text datasets. We choose to fuse the GCN-based model TextGCN and the transformer-based model pre-training Bert with each other.

Currently, GCN is used extensively in text classification tasks, such as link prediction[3], nodes classification[4]. Yao et al.[5] proposed that the textGCN model splits the document into words, a heterogeneous graph constructed by treating the words as nodes, and uses convolutional networks to obtain higher-order domain information. Petar et al.[6] proposed that the GAT model uses a convolutional neural network with masked self-attention, which is capable of processing graph-structured data. It has computational simplicity and allows for different weights of adjacency nodes. Haopeng et al.[7] proposed that the TG-transformer model partitions large-scale heterogeneous graphs into small batches of text graphs, learns effective node features for different text classifications, and improves global co-occurrence relations of words in the corpus. Tang et al.[8] proposed DGEDRT model which contains Bi-transformer and GCN. There is combining the flat representation learned from the transformer with the graphical representation learned from the corresponding dependency graph. The GCN approach allows the text to be represented as a graph, using the relationship between points and edges to obtain information about the features of the text.

The Bert model is a large pre-trained model proposed by Google. Currently combining Bert with GCN used extensively in NLP tasks. In this hybrid model, the Bert model is used to capture the global information of a sentence or document, and the GCN model with the convolutional network to complement the global information in the document. Lu et al.[9] proposed that VGCN-BERT model to construct a word co-occurrence-based lexical graph using graph convolutional networks, which uses word embeddings and graph embeddings together into the Bert to obtain local and global information. Qi et al.[10] proposed to apply the BertGCN in the Natural Language Inference. Co-dependent grammar trees are constructed from sentences which entered together into the GCN. Bert enhances the text representation in existing models. Shang et al.[11] proposed the G-BERT model that the first to bring the language model pre-training schema into the healthcare domain, in order to get more information from the data, G-BERT constructs an adaptive BERT model on the discarded single-visit data for visit representation.

Our work inspired by the above work and used the BertGCN hybrid model in Chinese medical short text classification. The disadvantages of short Chinese texts include short text sequences, incomplete utterance structure and noise in the semantic sequence. A node-based heterogeneous graph is constructed using a graph convolutional network, at the same time, we can obtain information about the points and edges in the graph structure to make a feature matrix. In the Bert model, we can get the global information using the mask mechanism. And then the global feature information pass through the GCN model. On the one hand, we can obtain higher-order matrix information by convolution which complement the global features. On the other hand, the node obtains local information which fuse the feature information of surrounding nodes and edges. We propose a new model BertGCN applied to the classification of Chinese medical short texts. Experimental results show superiority over other benchmark models.

3. METHOD

The BertGCN Model consists of two parts: Bert and GCN. In the Bert part, this processes the global information of points and documents in the heterogeneous graph. we use the output feature of the F_n^{bert} token as the document embedding, followed by a feedforward layer to derive the Bert final prediction. In the GCN part, this obtains the higher-order neighbourhood information of the text to prevent the weight enhancement of invalid nodes in the text. We use the output feature of the F_n^{gcn} token as the GCN final prediction. The Bert model is a Transformer-based bidirectional network structure, which use Mask increase the amount of information available in the field of view and use the Encoder part of the transformer as a basic unit of Bert. In this part, we use multi-head self attention to maximize the attention value of words in the text.

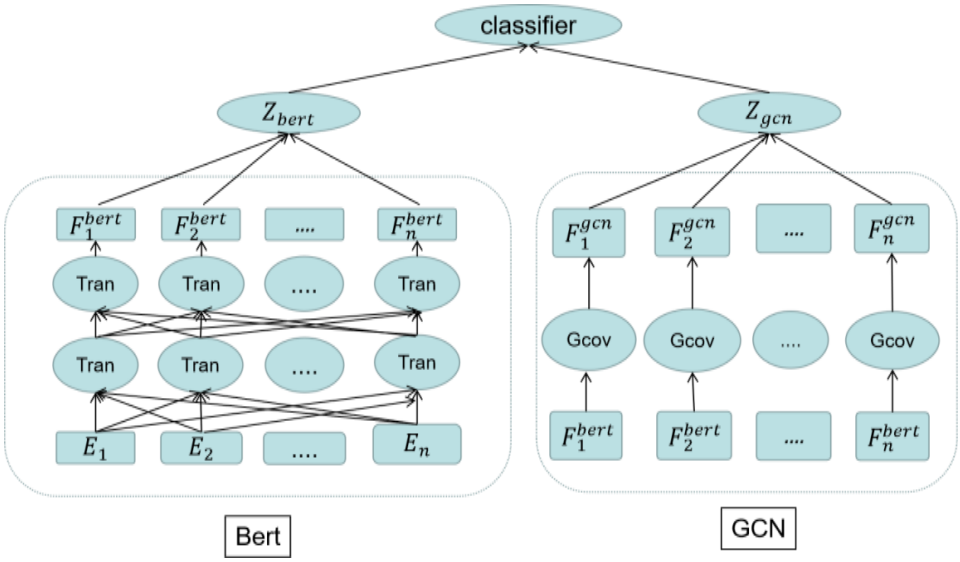

Figure 1. Architecture of BertGCN model.

As shown in Figure 1. E_1, E_2, E_n denote the input text initialization feature, which includes the information of the word embedding, position embedding and segment embedding. Assume that there are three input matrices $\mathrm{Q} \in R^{n\times \mathrm{d}_k}$, $\mathrm{K} \in R^{m\times \mathrm{d}_k}$, $\mathrm{V} \in R^{m\times \mathrm{d}_v}$ (n and m are the length of two inputs) after dotting with the text initialization feature, which represent the queries, keys and values respectively. QK^T gets the correlation between individual words of the text. $\mathrm{Q}' \in R^{n\times d_v}$ indicates the word attention value after normalized. d_v and d_k are the dimension size of values and keys respectively, at the same time $\mathrm{Q}' \in R^{n\times d_v}$ is the feature input information for GCN part. Z_{bert} denotes the possibility of selecting the maximum of these after regularization as the text classification.

$$
\begin{aligned}
Q' &= Softm\,\mathrm{ax}(\frac{QK^T}{\sqrt{d_k}})V \\
Z_{bert} &= soft\,\max(WQ') \\
Z_{\mathrm{gcn}} &= soft\,\max(\tilde{A}\,\mathrm{Re}\,lu(\tilde{A}Q'W_0)W_1)
\end{aligned}
\tag{1}
$$

The GCN part is made up of two layers Convolution network, $\widetilde{\mathrm{A}}\mathrm{Q}'W_0$ is the result of first layer of convolution to obtain the representation layer. $\widetilde{\mathrm{A}} = \mathrm{D}^{-\frac{1}{2}}A_0\mathrm{D}^{-\frac{1}{2}}$ represents the adjacency matrix by averaging and scaling its own vector information. D is the degree matrix for adjacency matrix $\widetilde{\mathrm{A}}$. $\widetilde{\mathrm{A}}\mathrm{ReLU}(\widetilde{\mathrm{A}}\mathrm{Q}'W_0)W_1$ is the result of second layer of convolution. W_0, W_1 as the weight matrix. After regularization, we can obtain the feature information in the heterogeneous graph. Z_{gcn} indicates the maximum possible classification of the current text after normalisation.

$$
\mathrm{Z} = mZ_{gcn} + (1-m)Z_{bert} \tag{2}
$$

Parameter m determines the proportion of Bert and GCN. $m \in (0,1)$ denotes the weights for the two models. m=0 means that the training model only uses Bert, m=1 means that the training model Bert only uses GCN. The output of the model is the maximum likelihood of the document in each classification, which is used to determine the label.

4. EXPERIMENTS

4.1 Data

As show in Table 1, the two datasets are all from the open source datasets in the network. The CMID[12] dataset divides the training data and the test data by 3:1. The amount of data in CHIP_CTC[13] is divided by the provider. The constructed large-scale heterogeneous graph consists of nodes and edges. The nodes contain word nodes and document nodes. For edge information, on the one hand, it comes from the relationship between words, on the other hand, it comes from the relationship between words and documents.

Table 1. The dataset information.

Dataset	Train	Test	Node	Edge
CMID	2883	22962	5681	316725
CHIP_CTC	961	7682	33183	1608597

4.2 Experiment result

As shown in Table 2, BertGCN showed the best results, compared to the other models. The experimental results are averaged over ten training sessions. In the training process, TextGCN is used for corpus preprocessing, the output matrix features are labelled usingCLS as the embedding of the document, and then regularisation is used to obtain the final prediction. From the result, we can see that compared with the BertGAT model the accuracy increased by 10.39% in the CHIP_CTC dataset. This is because that the CHIP_CTC dataset has a large amount of data compared to the CMID dataset.

In CMID dataset, BertGCN also has the best performance result compared to other models, it increased by1.21%. BertGAT model changes the weight of edges in a heterogeneous graph from fixed to compuTable weights, and the attention value is determined by the degree of influence that other nodes around the central node. The classification criteria in GAT are determined by the centre node information. In The GCN model, the attention value of the centre node is influenced by both the surrounding nodes and the edges. The higher matrix information of the central node is obtained by downsampling in the graph network.

Table 2. The accuracy of the model in different dataset (%).

Model	CMID	CHIP_CTC
TextGCN	57.60	61.57
GAT	59.79	67.22
Bert	60.73	57.55
BertGAT[14]	61.25	65.87
BertGCN	**62.46**	**76.26**

BertGCN performs better compared to the TextGCN model in CMID and CHIP_CTC dataset. The reason that TextGCN using the core mechanism of its graph convolutional network is an important method for building heterogeneous graphs. It can incorporate node and edge information in text initialisation features. However, large-scale pre-training models are able to obtain contextual global information. On the dataset we can fine-tuned to generate the most suiTable pre-training model. Therefore, the hybrid model BertGCN has advantageous compared with other benchmark models, In the model we use graph form to preserve data and deepen the model's understanding of the text during training in order to achieve improved model accuracy.

4.3 Experiment parameters

The TextGCN model with a two-layer graph convolution model perform best after testing. Multiple layers of GCNs can lead to poor differentiation of nodes and overfitting problems. In order to prevent overfitting and ignore a certain number

of neurons, the drop_out of the model is 0.5. The Bert model use Bert-base-Chinese as pretrained_bert, which has 8 attention heads. The learning rates in GCN and BERT are le-3 and le-5, respectively. The best performance is obtained by testing with a parameter m of 0.7.

5. CONCLUSION AND FUTURE WORK

In this paper we propose the use of BertGCN in Chinese medical short texts, which obtains the best performance capability in large-scale pre-trained models and inductive learning. We use GCN to construct a heterogeneous graph for the corpus, and then make the nodes and edges in the graph as document embedding information. This will enhance the information representation of textual information in the training. Finally, we use BertGCN to obtain local and global features information. The expressive power of BERTGCN was demonstrated from the experimental results. However, in our work, the small amount of data and incomplete node information resulted in a lack of textual information. In future research, text region is first performed for nodes in heterogeneous graphs, and the ability to classify text in specific domains is improved by obtaining additional features of the nodes in the knowledge graph of the feature domain.

REFERENCES

[1] Tayal, K., Nikhil, R., Agarwal, S., et al., "Short text classification using graph convolutional network," NIPS Workshop on Graph Representation Learning, (2019).

[2] Kateb, F. and Kalita, J., "Classifying short text in social media: Twitter as case study," International Journal of Computer Applications 111(9), 1-12 (2015).

[3] Mohamed, S. K., Nounu, A. and Nováček, V., "Biological applications of knowledge graph embedding models," Briefings in Bioinformatics 22(2), 1679-1693 (2021).

[4] Chen, J., Chen, S., Bai, M., et al., "Graph decoupling attention markov networks for semisupervised graph node classification," arXiv:2104.13718v2, (2022).

[5] Yao, L., Mao, C. and Luo, Y., "Graph convolutional networks for text classification," Proc. of the AAAI Conf. on Artificial Intelligence, 7370-7377 (2019).

[6] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P. and Bengio, Y., "Graph attention networks," arXiv preprint arXiv:1710.10903, (2017).

[7] Zhang, H. and Zhang, J., "Text graph transformer for document classification," Conf. on Empirical Methods in Natural Language Processing (EMNLP), (2020).

[8] Tang, H., Ji, D., Li, C., et al., "Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification," Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, 6578-6588 (2020).

[9] Lu, Z., Du, P. and Nie, J. Y., "VGCN-BERT: Augmenting BERT with graph embedding for text classification," European Conference on Information Retrieval, 369-382 (2020).

[10] He, Q., Wang, H. and Zhang, Y., "Enhancing generalization in natural language inference by syntax," Findings of the Association for Computational Linguistics: EMNLP 2020, 4973-4978 (2020).

[11] Shang, J., Ma, T., Xiao, C. and Sun, J., "Pre-training of graph augmented transformers for medication recommendation," Proc. of the Twenty-Eighth Inter. Joint Conf. on Artificial Intelligence, 5953-5959 (2019).

[12] Chen, N., Su, X., Liu, T., et al., "A benchmark dataset and case study for Chinese medical question intent classification," BMC Medical Informatics and Decision Making 20(3), 1-7 (2020).

[13] Zhang, N., Chen, M., Bi, Z., et al., "Cblue: A Chinese biomedical language understanding evaluation benchmark," arXiv preprint arXiv:2106.08087, (2021).

[14] Lin, Y., Meng, Y., Sun, X., et al., "Bertgcn: Transductive text classification by combining GCN and bert," arXiv preprint arXiv:2105.05727, (2021).