Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Research on Text Classification Based on BERT-BiGRU Model

To cite this article: Qing Yu *et al* 2021 *J. Phys.: Conf. Ser.* **1746** 012019

Research on Text Classification Based on BERT-BiGRU Model

Qing Yu, Ziyin Wang and Kaiwen Jiang*

Tianjin Key Laboratory of Intelligence, Computing and Network Security, Tianjin University of Technology, Tianjin, China

*Corresponding author email:15902224359@163.com

Abstract. Text classification is a typical application of natural language processing. At present, the most commonly used text classification method is deep learning. Meanwhile there are many difficulties in natural language processing, such as metaphor expression, semantic diversity and grammatical specificity. To solve these problems, this paper proposes the structure of BERT-BiGRU model. First, use the BERT model instead of the traditional word2vec model to represent the word vector, the word representation is calculated according to the context information, and it can be adjusted according to the meaning of word while the context information is fused. Secondly BiGRU model is attached to the BERT model, BiGRU model can extract the text information features from both directions at the same time. Multiple sets of experiments were set up and compared with the model proposed in this paper, according to the final experimental results, using the proposed BERT-BiGRU model for text classification, the final accuracy, recall and F1 score were all above 0.9. It shows that BERT-BiGRU model has good performance in the Chinese text classification task.

Keywords: Deep learning; Text classification; BERT; BiGRU model.

1. Introduction

With the wide popularization of social media, we are faced with a variety of news data. By identifying and effectively classifying these data, we can better understand what information netizens are more interested in and at the same time we will effectively supervise public opinion on the internet. Therefore, how to use natural language processing related techniques to analyze the type of news in internet has become one of the hotspots of current research [1].

Recently, more and more researchers have begun to use deep learning technology in the field of text sentiment classification[2]. The word vector is used to represent the text information, and the representation result is low dimensional and dense, so as to better represent the text. Then, neural networks such as CNN(Convolutional Neural Network) and RNN(Recurrent Neural Network) are used to automatically acquire feature expression. Compared with the traditional machine learning model, deep learning is more suitable for text classification tasks.

2. Correlational Research

At present, there have been many researches on the classification of texts.

In 2014, Kim[3] applied CNN to text classification task, which improves the accuracy of text classification to a certain extent. Wang et al[4] proposed a model of twitter text emotion analysis by using CNN. But CNN will ignore the context of the text. RNN[5] is capable of learning input of any length sequence, and mainly adopts Bidirectional Recurrent Neural Network (BiRNN) in order to learn the relationship between before and after sentence in the text. With the increase of input information, RNN is required to remember too much information, which leads to information redundancy and gradient disappearance[6]. Therefore, Hochreiter S et al[7] proposed long short-term memory (LSTM) to solve this problem. Huang et al[8] proposed a model of using BiLSTM to conduct emotion analysis of Chinese text. Due to the complex structure of LSTM, a new kind of recurrent neural network has been proposed based on LSTM, which is called gate controlled circulating neural network (GRU)[9]. In order to synthesize the context of the article, Cao et al[10] used BiGRU model to classify Chinese texts. This model is simpler, has fewer parameters, and has a faster convergence speed, which also shows a good effect in natural language processing tasks. After that, a large number of hybrid network models appear for classification.

In 2018, the pre-training model began to rise, making a major breakthrough in the field of NLP. ELMO[11], GPT, BERT[12] are successively appeared. Yu et al[13] proposed a method based on BERT-BiLSTM model to classify texts, and proved by experiments that compared with word2vec based BiLSTM model, the accuracy of BERT-BiLSTM model was higher.

For the purpose of achieve better classification result. We propose the model use of BERT to extract the feature representation of text. On this basis, we propose a bidirectional gate recurrent unit text classification model. This model makes the extracted text features more accurate, and we can prove the effectiveness of the proposed BERT-BiGRU model through comparative experiments.

3. BERT-BiGRU Model

The BERT-BiGRU model is shown in Figure1. It is divided into three parts:First, the semantic representation of each text is obtained through BERT model training, and the vector representation of words is obtained. Then the vector representation of the word is input into BiGRU model for further analysis and extraction of semantics. Finally connect the final word vector to softmax layer for text classification.

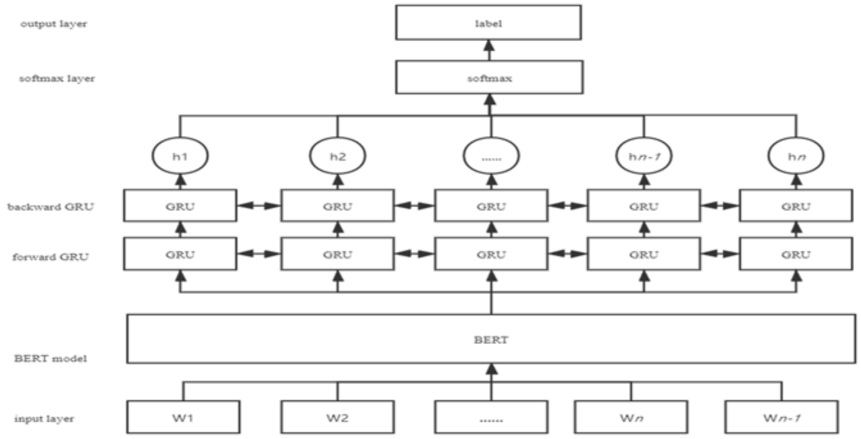

Figure 1. BERT-BiGRU model.

3.1. The Structure of BERT

The BERT model is the modification of Transformer structure. Transformer structure use the Encoder-Decoder structure. The BERT model remove the Decoder of Transformer only retain the Encoder of Transformer and denote as symbol *Trm.* The BERT model shown in Figure 2 is composed by multilayer bidirectional Encoder.

BERT is essentially a language generation model, and its goal is to generate a pre-training language model. Pre-training can be understood as training the model with a large amount of data, generating a general language model, and then fine-tuning the model for different downstream tasks. To train the model is equivalent to train the appropriate parameters so that the model can understand the semantics correctly. The BERT model uses two unsupervised approaches: Masked LM and Next Sentence Prediction. And combine the two approaches for pre-training.

3.2. The Structure of BiGRU Model

GRU is a variant of LSTM. Since RNN has a serious gradient disappearance problem when processing sequences, the perception of nodes in front becomes lower and lower as nodes get further and further back. In order to solve the gradient disappearance problem, the neural network of long and short time memory(LSTM) is proposed. In order to solve the problem of multiple parameters and long training time, a gate controlled circulation neural network (GRU) was proposed, is also suitable for processing sequential data, and can memorize the information of previous nodes through "gate" to solve the problem of gradient disappearance. In contrast to the LSTM, the GRU has only two gates, are update gate and reset gate respectively, so there are fewer parameters, which can achieve the same effect as LSTM while reducing the training time. The GRU model diagram is shown as Figure 3.

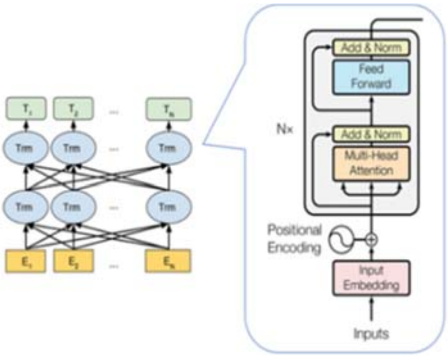

Figure 2. BERT.

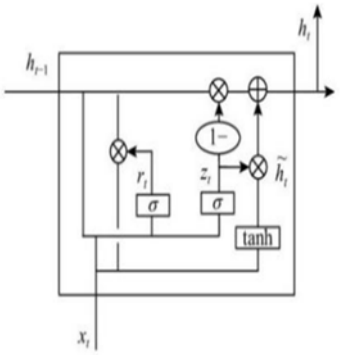

Figure 3. GRU structure.

In the following formal, x_t is the input data, $\mathrm{h_t}$ is the output of GRU, $\mathrm{r_t}$ is the reset gate and the z_t is update gate. z_t and r_t jointly control the calculation from h_{t-1} hidden state to h_t hidden state. The update gate controls both the current input data and the previous memory information h_{t-1}, and outputs a value z_t which value is between 0 and 1. z_t determines how much h_{t-1} is passed to the next state. The specific gate unit is calculated as:

$$z_t = \sigma(w_z \bullet [h_{t-1}, x_t]) \tag{1}$$

$$r_t = \sigma(w_r \bullet [h_{t-1}, x_t]) \tag{2}$$

$$\tilde{h}_t = \tanh(W \bullet [r_t \times h_{t-1}, x_t]) \tag{3}$$

$$h_t = (1\text{-z}_t) \times h_{t-1} + z_t \times \tilde{h}_t \tag{4}$$

In this formula, σ is the *Sigmoid* function, which changes the data to a value between 0 and 1 and acts as a gate control signal. GRU is a one-way neural network structure, the transmission of state is from the front to the back, but the current output state is related not only to the previous state but also to the subsequent state. In this case, bidirectional GRU is needed to solve this problem. BiGRU model perform information extraction in both directions, and the final output information is: $h_t^{(i)} = [\overrightarrow{h_t^{(1)}}, \overleftarrow{h_t^{(1)}}]$

$h_t^{(i)}$represents the BiGRU information for text *t*, $\overrightarrow{h_t^{(1)}}$represents the forward GRU information for text *t*, $\overleftarrow{h_t^{(1)}}$ represents the backward GRU information for text *t*, so the BiGRU model can obtain the information of the article both forward and backward. Moreover, it has the advantages of low complexity and short response time. The input of BiGRU model is the word vector obtained by BERT pre-training language model.

4. Experiment and Analysis

4.1. Experimental Data

To verify the validity of the proposed model. The THUCNews dataset is chosen. In this paper, 50000 pieces of data are selected as training set and 10000 pieces of data as test set. The data set is divided into:sports, finance, real estate, home, education, technology, fashion, politics, games, entertainment, 10 categories. The label is set to 0-9.

4.2. Comparative Experiments

This paper set up the following experiment:
(1)word2vec-BiGRU: The input text is represented by the word vectors which are obtained after the training of word2vec model, and these word vectors are used as the word embedding layer to access the BiGRU model for feature extraction and classification.
(2)ELMo-BiGRU: The input text is represented by the word vectors which are obtained after the training of ELMo model, and these word vectors are used as the word embedding layer to access the BiGRU model for feature extraction and classification.
(3)BERT-CNN: The corresponding representational word vectors are trained by BERT model for the input text, which are taken as input and classified by CNN neural network.
(4)BERT-RNN: The corresponding representational word vectors were trained by BERT model for the input text, which were then classified by RNN neural network.
(5)word2vec-RNN: This model is a traditional text classification model.

4.3. Model Parameter Setting

Table 1 is the parameter configuration of each model.

4.4. Evaluation Index

This paper adopts the evaluation standard applicable to text classification. Including precision(P), recall(R)and F1 score, the specific calculation formula is: P=TP/(TP+FP), R=TP/(TP+FN), F1=2PR/(P+R).

Table 1. Parameter of model.

model	word vector	batch size	learning rate	hidden layer_size
word2vec-BiGRU	300	8	3e-5	100
ELMo-BiGRU	768	4	1e-5	768
BERT-CNN	768	4	1e-5	768
BERT-RNN	768	4	1e-5	768
word2vec-RNN	300	8	5e-5	100
BERT-BiGRU	768	4	1e-5	768

4.5. Experimental Results and Analysis

The BERT-BiGRU model was used for training on the training set. The error and accuracy obtained are shown in Figure 4. As we have seen, the accuracy of the training set has achieved a good result. The final accuracy rate was 94.6%.

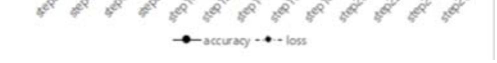

Figure 4. Experimental results.

The final results obtained by testing on the test set are shown in Table2. As its shown the *P*, *R* and F1 score of all categories reached more than 0.9, thus it can be seen that this model performs well in text classification. The comparison experimental results are shown in Table 3 and Figure 5. It can be seen that the effect of the designed comparative experimental model on text classification is not as good as that of the model proposed in this paper.

Table 2. Experimental results of Bert-BiGRU model.

	precision	recall	F1
sports	0.99	0.99	0.99
finance	0.93	0.98	0.96
real estate	1.00	1.00	1.00
home	0.91	0.83	0.87
eduction	0.94	0.96	0.95
techology	0.98	0.98	0.98
fashion	0.97	0.99	0.98
politics	0.95	0.95	0.95
games	0.89	0.99	0.94
entertainment	0.99	0.96	0.97
avg/total	0.95	0.95	0.95

Table 3. Comparison of experimental results table.

model	precision	recall	F1
word2vec-BGRU	0.66	0.79	0.72
ELMo-BiGRU	0.86	0.85	0.85
BERT-CNN	0.95	0.69	0.80
BERT-RNN	0.88	0.96	0.92
word2vec-RNN	0.77	0.71	0.74
BERT-BiGRU	0.95	0.95	0.95

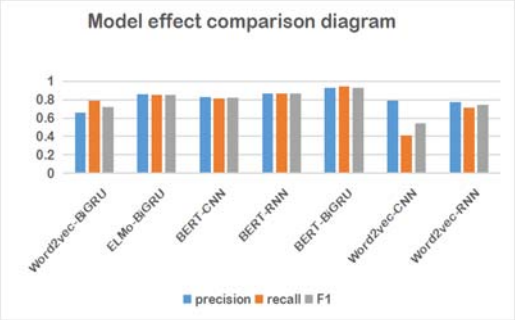

Figure 5. Comparison result of each model.

Compared with models word2vec-BiGRU and ELMO-BiGRU, we can verify the validity of BERT model. The word vector representation method of word2vec can not solve the polysemy problem in different environments. ELMo model is used to solve the polysemy of words. According to the experimental results, the performance of this model has been improved compared with word2vec to some extent, but the evaluation indexes are still inferior to BERT.

The comparison with model BERT-CNN and model BERT-RNN proves the validity of feature extraction and text classification based on BiGRU model. BERT model is adopted in the text representation of these models. While all of them are based on BERT model, BiGRU model has a better effect in obtaining the semantic features of text in both forward and backward aspects compared with CNN and RNN neural network.

By comparing with the traditional text classification model word2vec-RNN, the superiority of the new model is demonstrated. The final experiment shows that the text classification method based on BERT feature representation combined with BiGRU performs better on the data set.

5. Conclusion

Natural language processing is a hotspot in the field of artificial intelligence. The complex and changeable language features are the difficulties in this field. this paper proposes to use BERT model to obtain the feature representation of text, and input the obtained feature representation into the BiGRU model for further feature extraction so as to carry out more accurate text classification. The model presented in this paper has achieved good results in experiments with data sets.There are still some problems in this experiment. For example, BERT model is a depth model and have a large number of parameters, so training BERT model consumes a lot of hardware resources and time. This experiment can only rely on the pre-training model published by Google, we hope to optimize it in our future work.

References

[1] Nasukawa.T, Yi.J, *Sentiment analysis: Capturing favorability using natural language processing.*In Proceedings of the International Conference on Knowledge Capture, New York, NY, USA, October 2003, pp:23-25.

[2] Loai Aljerf, Mazen Aljurf, Improvements in the Ecological and Nutritional Aspects of Down's Syndrome. Preprints 2020, 2020050512. https://doi.org/10.21203/rs.3.rs-30313/v1

[3] Kim.Y, *Convolutional neural networks for sentence classification.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, State of Qatar, 2014, pp:1746-1751.

[4] Y.H.Wang, C.Y.Zhang, B.L.Zhao, *Emotional analysis of Twitter text in the context of convolutional neural network.* IEEE J. data acquisition and processin.,vol.33,no.05,pp:157-163,Dev.2018.

[5] L.W Liu ,X.Yu, *Circulating neural network (RNN) and its application.*IEEE J. Science and Technology World.,vol.32, no.6, PP:54-55,Nov 2019.

[6] Jing.Li, Gulcehre.Caglar, Peurifoy.John. Gated orthogonal recurrent units: on learning to forget. Neural Computation 31(4), 765-783.

[7] Hochreiter S, Schmidhuber J. *Long short memory.*IEEE J. Neural computation, vol.9, no.8, PP:1735-1780, Aug 1997.

[8] F.Huang, X.Y.Liu, G.F.Liu, X.Yang,*Emotion classification depth model based on word2vec and bidirectional LSTM.* IEEE J. Computer application research,vol:36, no:12, PP:3583-3587, Aug.2019.

[9] Dey R, Salemt F M. *Gate-variants of gated recurrent unit(GRU) neural networks.* IEEE 60th International Midwest Symposium on Circuits and Systems,Medford, MA, United States, 2017, PP:1597-1600.

[10] Y.Cao, T.R.Li, Z.Jia, *BGRU: A new method of Emotion analysis based on Chinese text.* IEEE J. Computer Science and Exploration, vol:13, no:6, pp:973-981, Feb 2019.

[11] Peters M, Neumann M, Lyyer M, *Deep Contextualized Word representations.* arXiv preprint arXiv:1802.05365 (2018).

[12] DEVLIN J, CHANG M W, LEE K, *Bert: Pretraining-of deep bidirectional transformers for language understanding.*IEEE J.Computation and Language. vol:23, no:2, pp:3-19, Dev 2008.

[13] Z.X.Yu, K.F.Hu, *Study on medical Information classification of Bert-Att -biLSTM Model.* IEEE J, Computer Age, vol:3, no:6, pp:1-4, Jan 2020.