Sentiment analysis based on BiGRU information enhancement

Xing Yin, Changhui Liu*, Xiaodong Fang
School of Computer Science and Technology, Wuhan Institute of Technology
21807010054@stu.wit.edu.cn

Abstract. In view of the existing sentiment analysis technology, model design is often a combination of different network models, which may not give full play to the advantages of the network. This paper proposes a data pre-training based on BERT, and input the obtained data into BiGRU neural network layer to enhance the features, and achieve the feature enhancement by stacking and repeated use. The input of the second layer is the result of the output of the first layer, and the input of the third layer is the result of the output of the second layer, and then it is calculated multiple times in different subspaces to learn relevant information. The experiment was conducted on the Twitter data set, and the results showed that in this idea, The two-layer BiGRU classification effect is better, reaching an accuracy of 82.63%, the convergence rate is faster, and it is better than other classification models, and the classification work can be completed well.

1. Introduction

This article aims at the English text, based on the BERT preprocessing method, and introduces the fusion BiGRU model. One is a single BiGRU, and the other is a two-layer BiGRU. The second layer input directly uses the unprocessed first layer output, enriching feature information and strengthening The feature learning ability of the model, thereby improving the accuracy of text sentiment analysis. The third is to introduce another layer on the second basis to further strengthen the feature learning ability. Using the NLP comprehensive evaluation indicators Accuracy value and F1 value on the data set used in the experiment, the effectiveness of the multi-layer BiGRU model in text classification has been proved, and the advantages and disadvantages of the multi-layer BiGRU are compared, and the most effective classification is selected. A good fusion model.

Mikolov et al.[1] proposed to apply RNN to the text classification task, so that RNN can fully learn the information between the front and back of the text, but the features of RNN that rely on learning for a long time will be prone to the problem of gradient dispersion. Graves et al.[2] improved the recurrent neural network and proposed the LSTM (Long Short-Term Memory) model. The LSTM unit controls the information transmission through the input gate, forget gate and output gate. GRU is another variant of RNN. It was proposed by Cho[3] and others in 2014. Compared with LSTM, GRU has a simpler structure, fewer parameters, and less training time. Cao Yu et al.[4] proposed a bidirectional threshold recurrent neural network (Bidirectional Gated Recurrent Unit, BiGRU) to perform text sentiment analysis. BiGRU considers the connection between the current word in the text and the following text, and makes up for the deficiency of GRU. So this article uses BiGRU as the object of the composite network. The attention mechanism has also shown significant effects in the field of natural language processing. In 2014, the Google team used a combination of recurrent neural network and attention mechanism to determine the attention part that should be processed according to the current input and the previous state. Then Bahdanau et al.[5] applied the attention mechanism to the machine translation task. It is its first application.
in natural language processing. Kokkinos et al\[6\] proposed the self-attention mechanism and successfully introduced the self-attention mechanism into the sentiment analysis task.

2. Model of BERT fusion BiGRU

2.1. BERT pre-trained language model

The first step of NLP work is to train the language model: convert the language and characters that exist in real life into vectors that can be calculated by the computer. The previous preprocessing methods were mainly Word2Vec and ELMo models. Word2Vec is the static pre-training technology proposed by Mikolov\[7-8\] in 2013. ELMo is a dynamic pre-training model proposed by Peters\[9\] and others. Both have played a significant role in the pre-processing link. However, both the Word2Vec and ELMo models have their own problems. From the Word2Vec to the ELMo model, the context-independent static vector becomes the context-dependent dynamic vector, and the ELMo model adopts bidirectional LSTM modeling to achieve contextual information reading.

In 2018, Devlin et al.\[10\] proposed the BERT pre-training language model. As formula (1):

\[
P(S) = p(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} p(w_i|w_1, w_2, \ldots, w_{i-1})
\]

The full name of BERT is Bidirectional Encoder Representations from Transformers. In order to better integrate the context of the left and right sides of the marked word, BERT uses a bidirectional Transformer as the encoder; compared to word2vec, BERT gradually moves the activities of downstream NLP tasks to pre-training to generate word vectors.

The most important part of BERT is the two-way Transformer coding structure. Transformer abandons the traditional RNN's cyclic network structure and completely models a piece of text based on the attention mechanism. Transformer coding unit as shown in Figure 1:

![Figure 1 Transformer Coding Unit](image)

Because BERT is a general pre-trained model, BERT has only encoder but no decoder. The main module of the coding unit is the self-attention part, as shown in formula (2):

\[
\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V
\]

Among them, Q, K, V are the input word vector matrix, dk is the input vector dimension. The core idea is that words in each lexical position can ignore the direction and distance, and directly encode each word in the sentence. Calculate the relationship between each word for all words in this sentence, and then use these relationships to adjust the weight of each word, and then you can get a new expression for each word. This new representation not only contains the word itself, but also the relationship between other words and the word, so it can express more comprehensive information than a simple word vector.
In order to strengthen the model's computing power for different positions and increase the effect of the attention unit, Transformer refers to the "multi-head" attention mode, such as formula (3-4):

\[
\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, ..., \text{head}_n)W^o
\]  

(3)

\[
\text{head}_i = \text{Attention}(QW^i, KW^i, VW^i)
\]  

(4)

At the same time, in order to solve the degradation problem in deep learning, the Transformer coding unit adds a normalization layer and a residual network, as shown in formula (5-6):

\[
\text{FFN} = \max(0, xW_1 + b_1)W_2 + b_2
\]  

(5)

\[
\text{LN}(x) = \alpha \frac{x_i - \mu}{\sigma^2 + \varepsilon} + \beta
\]  

(6)

In natural language processing, timing features are an important feature because Transformer has neither RNN recurrence nor CNN convolution. The Transformer model uses a sine wave similar to the periodic change of analog signals, such as formula (7-8), BERT The input is the sum of word embedding, position embedding, and type embedding.

\[
\text{PE}(\text{pos}, 2i) = \sin (\text{pos} / 10000^{2i/d_{\text{model}}})
\]  

(7)

\[
\text{PE}(\text{pos}, 2i + 1) = \cos (\text{pos} / 10000^{2i/d_{\text{model}}})
\]  

(8)

The BERT model further increases the generalization ability of the word vector model, fully describes the character characteristics of the character level, the word set, the sentence set and even the relationship between sentences, and the distributed expression of the word is better.

2.2. BiGRU layer

GRU (Gated Recurrent Unit) is a specialized recurrent neural network. GRU merges the forget gate and the input gate into a single update gate, and mixes the hidden state and the cell state at the same time. Its unit structure is shown in Figure 2:

![GRU Coding Unit](image)

Figure 2 GRU Coding Unit

The specific calculation process is as formula (9-12):

\[
Z_t = \sigma(W_z \cdot [h_{t-1}, x_t])
\]  

(9)

\[
r_t = \sigma(W_r \cdot [h_{t-1}, x_t])
\]  

(10)

\[
h_t = \tanh (W_c \cdot [r_t \cdot h_{t-1}, x_t])
\]  

(11)

\[
h_t = (1 - z_t) \cdot c_{t-1} + z_t \cdot \tilde{h_t}
\]  

(12)

Where \(\sigma\) is the sigmoid function and \(\cdot\) is the dot product. \(x_t\) is the input vector at time \(t\), \(h_t\) is the hidden state, and also the output vector, which contains all the valid information at the previous \(t\). \(z_t\) is an update gate, the control information flows into the next moment; \(r_t\) is a reset gate, the control information is lost; the two jointly determine the output of the hidden state.
2.3. *Fusion BiGRU model*

In order to fully utilize the BiGRU structure, the experiment uses 3 neural network structures to select effective features, as shown in Figure 3:

2.3.1. *BiGRU-1*

BiGRU-1 simply uses the BiGRU unit, as shown in Figure 3(a). The input of the BiGRU unit comes from the context window of words. For every word $C_t (1 \leq t \leq n)$:

$$x^t = v^{t-k_1}_c \oplus \cdots \oplus v^{t+k_2}_c$$  \hspace{1cm} (13)

![Figure 3 (a) Fusion BiGRU model](image)

The input $x_t$ of the BiGRU unit comes from the concatenation of $x^t((t-k_1):(t+k_2))$ word vectors, and $k_1$ and $k_2$ represent the number of words in the context of the word $c_t$, respectively. The output of the BiGRU unit is linearly converted and used as the final annotation process.

2.3.2. *BiGRU-2*

BiGRU-2 can stack BiGRU units layer by layer to form a multi-layer structure. Here, two layers are selected, that is, the output of the first layer is used as the input of the second layer without any transformation. The structure is shown in Figure 4(b).

![Figure 3 (b) Fusion BiGRU model](image)

2.3.3. *BiGRU-3*

BiGRU-3 is an extension of BiGRU-2. On the basis of BiGRU-2, another layer of BiGRU unit structure is stacked as shown in Figure 5(c).
3. Experiment

3.1. Experimental Data
This article uses the Twitter comment data set, which has sentiment tags. The tags are divided into two categories [0,1], positive emotion is 1, negative emotion is 0. And verify and analyze the sentiment analysis method of BERT-based fusion BiGRU model proposed in this paper. The data set settings are shown in Table 1.

| Table 1 Experimental Data Settings |
|-----------------------------------|
| Parameter | Value |
| Data      | 36000 |
| Validation set | 4000  |
| Test set  | 10000 |
| Total     | 50000 |

3.2. Parameter Settings
For the learning of the network, the setting of parameters is particularly important, it will directly affect the quality of the network. The specific settings of the model parameters in this paper are shown in Table 2.

| Table 2 Parameter Settings |
|----------------------------|
| Parameter          | Value |
| maxlen             | 128   |
| epochs             | 5     |
| BiGRU_dim          | 100   |
| Dropout            | 0.2   |
| lr                 | 1e-5  |
| Batch_size         | 32    |

3.3. Experimental Environment
The environment used in all tests is shown in Table 3.

| Table 3 Experimental Environment Configuration |
|-----------------------------------------------|
| Parameter      | Value   |
| Language       | Python3.6 |
| Tool           | Jupyter notebook |
| Frame          | Keras2.1.5 |
| Processor      | GPU     |

3.4. Experimental Procedure
In order to prove the effectiveness of the model, the above parameters were fixed during the test, and the comparison of the following tests was completed respectively:
1) CNN-BiGRU[11]: C-BG for short. Based on the Word2Vec pre-training model, first add a CNN network, and then add a BiGRU network.
2) CNN-BiLSTM-Attention\textsuperscript{[12]}, C-BL-A for short. Based on the Word2Vec pre-training model, first add a CNN network, then add a BiLSTM network, and finally introduce the Attention mechanism.

3) BERT: Single BERT network.

4) BiGRU-1: Simply use the BiGRU unit.

5) BiGRU-2: Stack BiGRU units layer by layer to form a multi-layer structure to enhance the features. Here, two layers are selected. The output of the first time is used as the input of the second layer. A dropout layer of size 0.2 is added between the two layers to reduce the training process fitting.

6) BiGRU-3: On the basis of BiGRU-2, stack a layer of BiGRU units to further enhance the features to form three layers, and add Dropout layer of size 0.2 between each layer.

3.5. Evaluation Index

The evaluation indicators in this article are Accuracy, Precision, Recall and F1 value. The specific formula is as (14-17):

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (15)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (16)
\]

\[
F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision + Recall}} \quad (17)
\]

TP is the number of positive emotions predicted as positive, FP is the number of negative emotions predicted as positive, TN is the number of negative emotions predicted as negative, and FN is the number of negative emotions predicted as positive.

3.6. Result Analysis

The experiment trains the model on the training set and scores it with the verification set, and finally calculates the result on the test set. The results of the five sets of models are shown in Table 4.

Table 4 Test results of models

|     | Accuracy | Precision | Recall | F1   |
|-----|----------|-----------|--------|------|
| C-BG | 0.736    | 0.7017    | 0.7246 | 0.7136 |
| C-BL-A | 0.7428  | 0.7344    | 0.6986 | 0.7161 |
| BERT | 0.8209   | 0.7536    | 0.8506 | 0.7991 |
| BiGRU-1 | 0.8221 | 0.7688    | 0.8226 | 0.7948 |
| BiGRU-2 | 0.8263  | 0.7767    | 0.8214 | 0.7984 |
| BiGRU-3 | 0.8149  | 0.75      | 0.8371 | 0.7912 |

Table 4 shows the comparison results of the 6 groups of models on the test set. From the two most intuitive comprehensive evaluation indicators Accuracy, F1. The use of the BERT pre-training model is better than the first two groups of models, which proves that the introduction of BERT does greatly improve the accuracy of sentiment analysis. The first and second sets of experiments also illustrate the necessity of the Attention mechanism. BERT integrates the Attention mechanism Which is also crucial. In the last three groups of fusion models, BiGRU can indeed enable the model to learn the forward and backward features of the serialized information more effectively, thereby optimizing the classification effect. In addition, through the comparison of Accuracy and F1 values between BiGRU-1, BiGRU-2 and BiGRU-3, as shown in Figure 6:
The result of BiGRU-2 is 0.4% higher than that of BiGRU-1, indicating that stacking BiGRU can indeed enhance the features, making the captured information more complete and conducive to classification. At the same time, BiGRU-3 is lower than the previous two. It is a lot, even 1% lower than the best model effect, BiGRU-2. This shows that the depth of the network model is too large, and it is easy to cause the model to overfit, which may not be conducive to the improvement of the results.

Convergence time. Similarly, by comparing the F1 values of the three fusion models, such as shown in Figure 7:

It can be found that the change curve of F1 value is also similar to that of Accuracy, and BiGRU-2 is better than BiGRU-1 than BiGRU-3. Taken together, the experimental results based on BERT preprocessing are better than the traditional results based on Word2vec preprocessing. In addition, the experiment effectively proves that the BiGRU model can capture the semantic combination of text more effectively, save longer context information, and further bring to the experiment Better results. And through the multi-layer BiGRU feature enhancement, the model has a better classification effect, in which two layers are the best stacking layers.

Use the model to predict the real single sentence, showing the real application function of the model. When selecting sentences, avoid the same sentences as those in the original data, otherwise the performance of the evaluation model will be affected. To facilitate understanding of the meaning of the output result, the value greater than 0.5 is set as positive, and the rest as negative. The single sentence test results are shown in Table 5.

| Forecast Result | Sentence | Sentiment |
|-----------------|----------|-----------|
| [[0.99151015]]  | squeaky out goin to bed talk to ya'll tomorrow night. | positive |
even the hoff can't bandaid the scar left by diversity winning and not the father son dancey team.

iremember wen the biggest deadline i had to worry about was the science fair project. i wanted to make the volcano.

conan is lookin good and his musical entrance was ridiculously long. like legit 2 min of applause.

Judging from Table 5, the prediction results of the BiGRU-2 model for randomly selected single sentences are correct, which more directly and clearly proves the successful application of the BiGRU-2 model in English sentiment analysis.

To sum up, for the data set used in this article, the fusion network is based on BERT preprocessing, and then the BiGRU-2 model is introduced to enhance the features obtained, which can more effectively improve the accuracy of text classification, and has good applications ability.

4. Conclusion
This paper proposes an English text sentiment analysis method that achieves feature enhancement through multi-layer BiGRU stacking. The multi-layer BiGRU stack is used to enhance the feature value to obtain more long-distance context information, while avoiding the gradient disappearance and gradient explosion problems caused by ordinary recurrent neural networks. And BERT preprocessing brings more convenience to text processing through its Transformer architecture and self-attention mechanism, which successfully improves the starting point of this project. Finally, it is concluded that the best classification effect can be obtained by introducing two-layer BiGRU based on BERT preprocessing. But because the network is more complicated, the model time cost is increased to a certain extent. In the future, it is the goal of the next step to study how to improve the classification accuracy and reduce the time cost.

Acknowledgments
Supported by the National Natural Science Foundation of China (61103136); Supported by the Education Innovation Program of Wuhan Institute of Technology (CX2019239)

References
[1] Mikolov T, Sutskever I, Chen Kai, et al. Distributed representations of words and phrases and their compositionality [J]. Neural Information Processing Systems, 2013: 3111-3119.
[2] Graves A. Supervised Sequence Labelling with Recurrent Neural Network [C]//Proc of International Conference on Learning Representations. 2015:1-15.
[3] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [C]//Proc of International Conference on Learning Representations. 2015:1-15.
[4] Kokkinos F, Potamianos A. Structural Attention Neural Networks for improved sentiment analysis[C]//Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017:586-591.
[5] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
[6] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
[9] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.

[10] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[11] Miao Yalin, Ji Yichun, Zhang Shun, Cheng Wenfang, Peng Erlou. Application of CNN-BiGRU model in sentiment analysis of Chinese short text[J]. Information Science,,:1-6.

[12] Wang Liya, Liu Changhui, Cai Dunbo, Zhao Tongzhou, Wang Meng. Text sentiment analysis based on CNN-BiLSTM network with attention model[J].Journal of Wuhan Institute of Technology,2019,41(04):386-391.