A Text Category Detection and Information Extraction Algorithm with Deep Learning

Xiaohan Wu1,*, Zejun Wu2, Yuqi Feng3

1BNU-HKBU United International College (UIC), Division of Business and Management (DBM), China
2School of Cyber Science and Engineering, Wuhan University, Wuhan, China
3College of Mechanical and Electrical Engineering, Northeast Forestry University, Harbin, Heilongjiang, China

*Corresponding author: xiaohanwu123@bnu.edu.cn

Abstract. In order to solve the problem that the text classification model based on neural network is easy to over-fit and ignore the key words in sentences in the training process, a Bi-GRU Chinese text classification model based on hierarchical Attention mechanism is proposed. The model introduces the idea of layering, uses bi-directional gated cyclic neural network to learn the text representation at word level and sentence level, uses Self-Attention hierarchical model to obtain the information of the influence of words and sentences on text classification, shares the weight between embedded layer and softmax layer by binding, and uses AMSBound optimization method to obtain the optimal weight matrix quickly and effectively while reducing the parameters in the model. Two commonly used Chinese data sets, FudanSet and THUCNews, are tested on the long Chinese text classification data set FudanSet. The experimental results show that the accuracy, recall rate and F-score of this model are better than Text-CNN model, Attention-BiLSTM model and Bi-GRU_CNN model, and the accuracy, recall rate and F-score index are improved by 5.9%, 5.8% and 4.6%, respectively.

Keywords: Chinese text classification; bi-directional gated loop unit; hierarchical attention mechanism; weight binding; adaptive boundary gradient optimization method.

1. Introduction
With the arrival of the era of big data, the task of natural language processing is becoming more and more heavy. A large number of researchers have devoted themselves to the research of natural language processing, among which text classification has attracted more and more attention[1].

With the continuous development of deep learning, researchers have introduced it into the field of natural language processing, and then used in text classification, emotional analysis and other tasks. For example, Text-CNN (text convolution neural network) model, improved attention short text classification method based on bidirectional long-term memory network, DC-BiGRU_CNN (convolution network based on dense link cyclic gated unit convolution network) for short text
classification, all of these are used to deal with short English text classification. However, due to the ambiguity of semantics, the complexity of part of speech and the asymmetrical distribution of various parts of information in the document, the model of dealing with English text can not be adopted directly in Chinese text. Then how to classify Chinese text more accurately and efficiently has become a research hotspot. As we all know, the contribution of each word in the document to the classification task is different, so it is very important to obtain the degree of contribution to improve the accuracy and efficiency of the text classification task [2].

To solve the above problems, this paper proposes a Chinese text classification model of bi-directional gated loop unit (Bi GRU) based on hierarchical Attention (attention) mechanism. This model mainly uses the GRU model to learn the representation of words and sentences. On this basis, the hierarchical Self-attention (self-attention) mechanism is used to give different weights to different words and sentences, and the AMSGrad optimization method is used to obtain the optimal weight matrix quickly and effectively, so as to improve the accuracy of classification [3].

2. BiGRU text Classification Model based on hierarchical Attention Mechanism

![Figure 1](image)

**Figure 1** Attention-BiGRU network structure diagram

The basic idea of this model is to deal with the hierarchical idea that documents are composed of sentences and sentences are composed of words. First, the word sequence is represented by two-way GRU, and then the Self-Attention mechanism is used to extract the importance of different words in different sentences, that is, the weight of each word. Then, on this basis, sentences are represented by
bi-directional GRU, and then the amount of information of different sentences in the document is extracted by Self-Attention mechanism. Finally, the document is vectorized, and then the document is classified by the softmax layer [4]. The network structure is shown in figure 1.

3. Experimental comparison and analysis

3.1. Data set
In order to prove the effectiveness of this model, two classical Chinese text classification data sets, FuDanSets and THUCNews, are selected in the experiment. FuDanSets is a Chinese text classification data set created by the Natural language processing Group of the International Database Center of the Department of computer Information and Technology of Fudan University, which is divided into 20 categories, including art, education and energy. The FuDanSets is divided into two parts: the training set and the test set according to 1:1.

THUCNews, launched by the Natural language processing Laboratory of Tsinghua University, is based on historical data from Sina News. There are 740000 news articles available. It is divided into 14 categories, including finance, lottery and sports. For the dataset THUCNews, 3000 training samples and 500 test samples are randomly selected for one category. The details of the dataset are shown in Table 1.

3.2. Model training
In order to prevent the GRU model from overfitting and gradient explosion, the following improvements are made.

In the process of model training, the Dropconnect method is applied to the hidden-to-hidden weight matrix \((W_{r}, U_{r}, W_{h}, U_{h}, W_{z}, U_{z})\), and the probability of the input weight of each node connected to it is changed to 0. At the same time, bind the weight We of the shared embedded layer and the weight Wc of the softmax layer to reduce a large number of parameters in the model. In addition, the Adam optimization method used in the later stage of the learning rate is too low, resulting in poor convergence. By analyzing the advantages and disadvantages of the commonly used Adam and SGD optimization methods, in order to solve this problem, using Adam, in the later stage to replace SGD, can effectively make use of the advantages of both and improve the convergence speed of the algorithm. AMSBound combines this to realize intelligence, gives the method of choosing the time to switch SGD, and adopts the learning rate of dynamic boundary to realize the smooth transition to SGD. This method not only keeps the advantages of fast initialization and hyper-parameter insensitivity of the adaptive method, but also shows good results on several standard benchmarks. Therefore, the AMSBound optimization method is used as the optimizer in model training to accelerate model training.

3.3. Experimental results and analysis
In the experiment, three indicators commonly used to measure the effectiveness of the neural network model, namely, accuracy, recall and F-score, are selected to verify the effectiveness of the proposed model.

The FudanSet dataset is compared with Text-CNN model, Attention-BiLSTM model and Bi-GRU_CNN model respectively, and the experimental results are shown in Table 1. From the experimental results, we can see that all the indicators of the model proposed in this paper have been improved. Compared with the better model, the classification accuracy has been improved by 5.9%, and the recall rate has increased by 4.6%, and the recall rate has increased by 5.8%.
Table 1: Experimental results of Fudan Set dataset

| Model                  | Recall rate /% | F-score /% | Accuracy /% |
|------------------------|----------------|------------|-------------|
| Text-CNN               | 82.1           | 82.4       | 82.8        |
| Attention—BiLSTM       | 82.9           | 83.1       | 83.3        |
| Bi—GRU_CNN             | 84.1           | 83.6       | 84.2        |
| Model in this paper    | 89.7           | 88.2       | 90.1        |

Table 2: Experimental results of THUCNews dataset

| Model                  | Recall rate /% | F-score /% | Accuracy /% |
|------------------------|----------------|------------|-------------|
| Text-CNN               | 82.1           | 82.4       | 82.8        |
| Attention—BiLSTM       | 82.9           | 83.1       | 83.3        |
| Bi—GRU_CNN             | 84.1           | 83.6       | 84.2        |
| Model in this paper    | 89.7           | 88.2       | 90.1        |

The THUCNews dataset is compared with Text-CNN model and Bi-GRU_CNN model respectively, and the experimental results are shown in Table 3. From the experimental results, we can see that all the indicators of the model proposed in this paper have been improved. Compared with the better model, the classification accuracy has been improved by 1.9%, and the recall rate has increased by 1.1%, and the recall rate has increased by 1.1%.

As you can see from Table 2, the number of documents in the THUCNews dataset is more than twice that of the FudanSet dataset, while the average length of the document is 1 prime 5 of the FudanSet dataset, and the maximum sentence length is 1 canary 4 of the FudanSet dataset, which indicates that the FudanSet dataset is a long text document, while the THUCNews dataset is a shorter document. Combining table 2 and table 3, we can see that the model proposed in this paper is more suitable for long Chinese document classification, although it is better than the current optimal model in terms of various indexes, but the improvement in FudanSet data set is greater.

Table 3: Time to train the model

| Model                  | Training model time / min |
|------------------------|---------------------------|
|                        | Fudan Set | THUCNews |
| Text-CNN               | 60         | 120      |
| Attention—BiLSTM       | 94         | 146      |
| Bi-GRU_CNN             | 92         | 200      |
| Model in this paper    | 74         | 128      |

When evaluating the performance of a model, in addition to comparing the accuracy of the target task, there is also an important performance index, that is, the efficiency of the model on the same configuration platform, that is, the time required to train the model. Therefore, finally, comparing the time of the training model, as shown in Table 4, we can see that the convergence speed of the proposed model is faster.

4. Conclusion

Text classification is one of the important tasks in natural language processing. In recent years, with the arrival of the era of artificial intelligence, the demand for the accuracy of long text classification is getting higher and higher. This paper first uses the forward and reverse learning of GRU cyclic neural
network which is suitable for dealing with long text sequence to represent the word vector of long text, then obtains the importance of the word in the sentence through Self-Attention, gives the corresponding weight to the word vector, and inputs it into the GRU network for forward and reverse learning to get the representation of the sentence, and then obtains the importance of the sentence in the document through Self-Attention. And give the corresponding weight to the sentence vector, get the representation of the document, and finally classify the document through the softmax layer. In the whole training process, the vector dimension is large, which leads to the efficiency not reaching the ideal state. In the next step, we will use the commonly used effective dimensionality reduction methods, such as convolution neural network, to obtain its keyword feature vector to improve the efficiency of the model.

Reference
[1] Zhang Xiaochuan, Yu Linfeng, Zhang Yihao. Short text similarity calculation based on LDA multi-feature fusion [J]. Computer Science, 2018 (9): 266–270.
[2] Ju Yaya, Yang Lu, Yan Jianfeng. LDA algorithm based on dynamic weight [J]. Computer Science, 2019 (8): 260–265.
[3] Chang Dong Dong, Yan Jianfeng, Yang Lu, et al. Topic model based on sliding window [J]. Computer Science, 2016 (12): 101–107.
[4] Liang Jiye, Qiao Jie, Cao Fuyuan, et al. Distributed representation Model for short text Analysis [J]. Computer Research and Development, 2018 (8): 1631–1636.
[5] Liang Jiye, Qiao Jie, Cao Fuyuan, et al. Distributed representation Model for short text Analysis [J]. Computer Research and Development, 2018 (8): 1631–1636.