News text classification based on hybrid model of Bidirectional Encoder Representation from Transformers and Convolutional Neural Network

Fan Wei¹,a, Fan Li²*
¹Department of Computer Science, Xi’an Shiyou University, Xi’an, China
²Department of Computer Science, Xi’an Shiyou University, Xi’an, China
¹Email: fwei@xsyu.edu.cn
²Corresponding author’s e-mail: 19212060600@stumail.xsyu.edu.cn

Abstract. At present, with the rapid development of Internet technology, text data show massive characteristics. Network media and new media platforms have become an important part of journalism and communication, and network news has become one of the important sources of information. In order to meet the needs of online news reading users and improve the efficiency of personalized news recommendation by content distribution platform, it is urgent to effectively manage and utilize online news. Based on this, this paper studies and implements the algorithm for news text classification, and the goal is to build a classification model for news text with higher classification accuracy. In this paper, BERT and CNN algorithm are combined to classify news texts. The core idea is to send BERT as the embedding layer into the CNN model. The experimental results show that this method is better than the simple BERT model or the simple CNN model.

1. Introduce

1.1 Research background
In today’s Internet, text information, as a relatively direct form of media, accounts for an increasing proportion. The main characteristics of these online text data are the development of high real-time and high complexity Internet, which makes countless users use these information data at all times. How to use text processing technology to extract valuable features of these information data and analyze the interests and concerns of each user has become the main research direction of the current Internet era⁰¹.

1.2 Research significance
As one of the most critical technologies in text information processing, text classification is currently mainly used in information retrieval, information supervision and knowledge mining. At present, the main classification models used in text classification are CNN, RNN, K-nearest neighbor algorithm, naive Bayes, support vector machine, Back Propagation neural network, BERT and other classification models, and they have achieved good results⁰². However, a single model often has its own defects. After training to a certain number of times, the accuracy of text classification does not rise or fall. Therefore, usually in the text classification, will improve a model or use a combination of
several algorithms to build a model for text classification\textsuperscript{[3]}. In this paper, news texts are classified by the model combining BERT and CNN algorithm. The model combines the advantages of the two algorithms and uses a multi-layer nonlinear mapping structure, thus effectively overcoming various defects in their own model.

2. Materials and Methods

2.1 BERT

BERT is based on bidirectional transformer, that is, the traditional meaning is bidirectional similar to ELMo model\textsuperscript{[4]}. Its structure is shown in Figure 1. According to the context, each word will be represented by left-to-right and right-to-left, and then concat will be used as the representation of the word in the future, and then the downstream task operation will be carried out.

Two common versions of BERT:

\[ \text{BERT}_{\text{BASE}}: L = 12, H = 768, A = 12, \text{TotalParameters} = 110M \]  \hspace{1cm} (1)

\[ \text{BERT}_{\text{LARGE}}: L = 24, H = 1024, A = 16, \text{TotalParameters} = 340M \]  \hspace{1cm} (2)

where \( L \) represents the number of layers of transformer, \( H \) represents the dimension of output, and \( A \) represents the number of multi-head attention.

2.2 CNN

CNN network has a total of five hierarchical structures: input layer, convolution layer, activation layer, pooling layer, and fully connected FC layer. CNN structure is shown in Figure 2.

Input layer: CNN model needs to preprocess the input\textsuperscript{[5]}. Convolution layer: This layer uses convolution kernel for local perception. Through local perception characteristics, the calculation parameters of the model are greatly reduced. Excitation layer: Excitation is actually a nonlinear mapping of the output of the convolution layer. Pooling layer: it is mainly used for feature dimension reduction, compression of the number of data and parameters and improvement of fault tolerance of the model\textsuperscript{[6]}. This paper uses maximum pooling, as shown in Figure 3. Output layer: dropout
operation is introduced in this layer to randomly delete some neurons in the neural network to solve the problem of possible overfitting.

3. Experimental design

3.1 Experimental environment of news text classification

The experimental environment is shown in Table 1.

| Experimental environment | Environment configuration |
|--------------------------|---------------------------|
| Device                   | PC                        |
| Operating system         | Windows 10                |
| CPU/Memory               | 4 core/16G                |
| Programming language     | Python                    |
| Deep learning framework  | Pytorch 1.9.0             |

3.2 data preparation

200,000 news headlines were extracted from THUCNews and the text length was between 20 and 30. There are 10 categories, namely finance, real estate, stocks, education, technology, society, current affairs, sports, games, and entertainment. 20,000 pieces of data for each category. The acquired data are relatively balanced, which avoids the impact of data imbalance on the experiment. The data set used in this article is divided into training set, validation set and test set, which respectively account for 180,000 pieces of data, 10,000 pieces of data, and 10,000 pieces of data.

3.3 experiment process

Firstly, the processed data is input into the BERT model, and then the final token-level output of the BERT model is obtained as the embedding as the input of CNN. Three different convolution kernels are used for convolution. The sizes of three different convolution kernels are [2, 3, 4], and the number of each convolution kernel is 256. The three feature maps obtained after convolution are maximum pooled, and then spliced and fully connected. Finally, softmax is used for normalization, and the class corresponding to the largest number of categories is used as the final prediction.

3.4 Experimental parameter setting

The main parameter settings of this experimental model are shown in Table 2.

| Parameter                  | Set value |
|----------------------------|-----------|
| convolutional kernel size  | (2,3,4)   |
| Number of convolution kernels | 256       |
| Batch size                 | 128       |
| Epoch                      | 3         |
| Dropout rate               | 0.1       |
3.5 Evaluation index

The precision (P), recall (R), F1 value and accuracy of the classification model used in the experiment are used as the evaluation indexes of the model. The calculation formula (3) - (6) shows that:

\[
P = \frac{TP}{TP + FP} \quad (3)
\]

\[
R = \frac{TP}{TP + FN} \quad (4)
\]

\[
F1 = \frac{2 \times P \times R}{P + R} \quad (5)
\]

\[
Acc = \frac{(TP + TN)}{(TP + TN + FN + FP)} \quad (6)
\]

where TP is True Positive; TN is True Negative; FN is False Negative; FP is False Positive.

4. Conclusions

The experimental results of news text classification are shown in table 3. As shown in Table 3, we can see that the accuracy of the three models is 0.9116, 0.9321 and 0.9471, respectively. The classification effect of using BERT sequence output as the embedding layer of tokens into CNN model is significantly better than that of BERT model and CNN model alone. Because this model combines CNN and BERT models at the same time, it can effectively obtain context semantic information, remove redundant information and obtain local features. To some extent, it overcomes the defects of BERT and CNN model and realizes complementary advantages in the model. News text classification has a wide range of applications. This paper proposes a hybrid model based on BERT and CNN to classify news text, which provides the accuracy of news text classification and reduces the classification error rate. The research in this article can also provide a certain reference for other researchers.

Of course, the model in the text still has certain limitations and is not suitable for the classification of long texts (more than 30 Chinese characters). In future work, the model can strengthen the classification effect of long texts.

Table 3. Classification results of three models.

| Model       | Precision | Recall | F1-score | Accuracy |
|-------------|-----------|--------|----------|----------|
| CNN         | 0.9121    | 0.9116 | 0.9117   | 0.9116   |
| BERT        | 0.9324    | 0.9321 | 0.9321   | 0.9321   |
| BERT-CNN    | 0.9475    | 0.9471 | 0.9471   | 0.9471   |

References

[1] Vandana K, Namrata M. (2012) Text Classification and Classifies:A Survey. International Journal of Artificial Intelligence & Applications (IJAIA), No.2:86-99.

[2] Zhang Y, Wallance B.A. (2015) Sensitivity Analysis of Convolutional Neural Networks for Sentence Classification. Computer Science: 162-166.

[3] Robert E.S, Yoram S. (2013) A Boosting-based system for text categorization. Machine learning 39:5-11.

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

[5] Yang Xiang, Hong Song. (2020) Application of CNN-ATTBiLSTM Hybrid Algorithm in Natural Language Processing. Journal of Vibration Testing and System Dynamics, 167-173.

[6] Wang, P., Xu, B., Xu, J., Tian, G., Liu, C., and Hao, H. (2016c) Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. Neurocomputing, 174(PB): 806-814.