Improved Convolutional Neural Network for Biomedical Text Classification

Zhenzhen Ma1*

1School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, 400000, China

*Corresponding author’s e-mail: s180231501@stu.cqupt.edu.cn

Abstract. In recent years, the combination of biomedical field and computer field is booming. Obtaining useful information from a large amount of biomedical text information is a research topic of great significance. Convolutional neural network has a good ability to extract useful features, so it is widely used in the field of text classification. In this paper, a novel approach for biomedical text classification based on improved convolutional neural network is proposed to solve the problem that deep convolutional neural network has a large amount of computation and can not perceive the relationship between levels well. In this paper, we use the combination of deep separable convolution and void convolution to improve the convolutional neural network. At the same time, we use the attention mechanism to classify biomedical literature. In addition, focusing loss function is used to improve the imbalance of biomedical texts. Experimental results show that the classification model in this paper is effective for biomedical texts.

1. Introduction

In recent foreign studies on text classification through deep learning, Ghiassi M et al. proposed to introduce dynamic artificial neural network algorithm to solve the problem of document classification[1]. Yoon Kim in 2014 will be introduced to the depth study of natural language processing field, and put the sentences into a two-dimensional matrix, the input to the CNN model training, through several groups of experiments to verify the rationality of the algorithm is combined with natural language processing and effectiveness, to a great extent improve the classification accuracy of text corpus[2]. However, this model still has some defects, and its running speed needs to be improved. The parameter setting of CNN model relies more on the experience value of users, resulting in a waste of time. Mingbo Ma proposed to use the N-gram model to explore the correlation between words, and effectively combine the model with deep learning to solve the problem of long-distance dependency between text contexts and improve the accuracy of text classification[3]. Rie Johnson used the one-dimensional structure of text (Word Order) as the input of CNN model, and at the same time, bag-of-word conversion was adopted in the convolution layer to complete Word variants, and the effectiveness of the algorithm was verified through experiments[4]. Siwei classify text by introducing RNN, capture context information through cyclic structure, reduce noise, automatically judge the importance of words in text through a maximum sharing layer, and improve text classification efficiency[5]. Ji Young Lee combined the advantages of CNN and RNN models and proposed a C_LSTM algorithm, which extracted a series of high-level phrase representations by CNN and then transferred them to LSTM for classification. The experiment proved that the C_LSTM algorithm could capture local features of phrases as well as global and temporal sentence semantics[6]. Zhou C proposed to combine CNN and RNN models to train text to obtain text vectors, and used ANN classifier to test the training text vectors, and the results proved...
that its classification accuracy could reach more than 90%[7]. Cao used CNN to learn the feature vector representation corresponding to each sentence and input the learned vector representation as the feature into the SVM classifier for sentiment classification[8].

Similarly, text classification technology is also widely used in the biomedical field. Ruihua Fang et al. used support vector machine to develop an automatic text classification system that could identify 10 specific types of bioscience literatures from massive literatures, including antibody, phenotype, breakout allele sequence, etc.[9]. K.B rettonnel Cohen, text mining is studied in the field of conversion of bioinformatics (the branch of natural language processing) the application of clinical researchers to determine through text mining technology basic biological research and clinical trials data corresponding to each other, and through the screening specific types of literature in preparation for the experimental study meta-analysis[10]. The most direct way to improve the performance of convolutional neural network is to increase the number of layers and the number of channels in each layer. However, as the depth and width of the network continue to increase, the parameters that the network needs to learn and the amount of computation continue to increase, overfitting is also likely to occur. CNN can capture local features very well, but it can't capture the global connection of text. At the same time, there is a problem of unbalanced data in biomedical texts, which will exert extreme influence on the classification results.

Therefore, this paper uses the combination of void convolution and deep separable convolution to improve the convolutional neural network, so as to obtain deeper features while reducing the amount of computation. Secondly, the attention mechanism is used to capture the global relationship of words, focusing more on the key words. Then the focused loss function is used to alleviate the data imbalance in biomedical texts. Finally, support vector machine algorithm is used to replace the traditional convolutional neural network softmax layer for text classification.

2. Materials and Methods

This paper classifies the biomedical literature. Firstly, the text is preprocessed and the word vector is represented. Secondly, the improved convolutional neural network is used for feature extraction. Then the multi-head self-attention mechanism is used to extract the key words and capture the global relationship. Finally, the support vector machine is used to classify the extracted text features and output the final results. The experimental process of this paper is shown in Figure 1.

![Flowchart of Biomedical Text Classification](image)

Figure 1 Flowchart of Biomedical Text Classification

2.1. Text Pre-Processing

Because most of the biomedical texts involved are SCI papers and other literatures, there are titles and clinical trials in the titles and abstracts. Therefore, in the text preprocessing stage, in addition to word segmentation and removal of stop words, we also need to carry out abbreviation expansion, word form restoration and case conversion operations. In this way, the preprocessed text can better express the text features. The NLTK module, a natural language processing tool set in Python, is used to preprocess text data. For preprocessed data, for a certain category, 0 in the first column means that it does not belong to a certain category, and 1 means that it belongs to a certain category. The second column is the processed title and content.

2.2. Word embedding

The computer can only recognize mechanical language. In order to enable the computer to recognize the text language, it is necessary to convert the words in the text into digital representations, that is, generate word vectors from the words and represent them as a data matrix.
The word2vec word vector training model mainly includes Continuous Bag-of-Words Mode (CBOW) and Skip-Gram model. The CBOW model predicts the current word by the words above and below the current word, while the Skip-Gram model (Figure 2 Shown). The words around the current word can be predicted by the current prediction word.

![Figure 2 Skip-Gram word vector training model](image)

This paper uses the Skip-Gram model to represent the preprocessed text data as word vectors, and generate word vectors of the form \( w = (d_1, d_2, \cdots, d_n) \). Among them, \( w \) represents the feature word, and \( d_i (0 \leq i \leq n) \) represents the i-th dimension of the feature word \( w \). The dimension of word vector trained in this paper is 128 dimensions, i.e. \( i=128 \). If there is a set of vocabulary \( \{w_1, w_2, \cdots, w_n\} \), the goal of the Skip-Gram model is to maximize \( J \):

\[
J = \frac{1}{N} \sum_{n=1}^{N} \sum_{-c \leq l \leq c, j \neq 0} \log p(w_{n+i} \mid w_n)
\]

Where \( n \) is the number of words in the vocabulary, \( i \) is the size of the context window, \( w_n \) is the current word.

After word2vec word vector training, the word vector representation of each word is obtained, and the complete information of the document cannot be obtained. Assuming that a document can be composed of multiple sentences, it can be expressed as \( d = \{s_1, s_2, \cdots, s_j\} \), and \( s_j \) represents the j-th sentence in the document. A sentence can be composed of multiple words, then the sentence can be expressed as \( s_j = \{x_1, x_2, \cdots, x_n\} \), \( x_j \) represents the i-th word in the sentence, and the word can be expressed as \( x_i = \{k_1, k_2, \cdots, k_m\} \), \( k_i \) represents the weight on the i-th dimension of the word. After the splicing of word vectors, the sentence can be expressed as:

\[
s = x_1 \oplus x_2 \oplus \cdots \oplus x_n
\]

The document is composed of multiple sentences, the document is expressed as:

\[
d = s_1 \oplus s_2 \oplus \cdots \oplus s_n
\]

Among them, \( \oplus \) represents the connector, and the data matrix is used as the data of the convolutional layer for feature extraction.

2.3. Improved-CNN

In the text classification task, the traditional CNN feature extraction method is limited by the convolution kernel. In order to obtain more features, it is often adopted to enlarge the convolution kernel to increase the number of receptive fields or stack convolution layers to obtain deeper features. However, it will face such problems as model parameter doubling, large amount of calculation, bloated model and gradient explosion easily. In order to alleviate the above problems, a combination of void convolution
and deep separable convolution is adopted to replace the ordinary convolution in this chapter, so that the convolutional neural network can expand the receptive field while reducing the computational load and obtaining multi-scale features. The structure of deep separable void convolution is shown in Fig. 3.

![Deep separable void convolution](image)

In Fig. 3, \( r \) represents void rate. Replace \( k \times k \) standard convolution with a deep convolution with a convolution window size of \( k \times k \) and then add \( 1 \times 1 \) standard convolution, which can effectively reduce the number of parameters in the convolution kernel of the same size and reduce the amount of calculation. At the same time, in order to make the small size convolution kernel obtain the large receptive field, the deep convolution of \( k \times k \) is changed into the deep void convolution with a void rate of \( r \). Therefore, the calculation formula of void convolution receptive field is as follows:

\[
F_j = F_{i-1} + (k_j - 1) \times r \times \prod_{n=1}^{i-1} s_n
\]  

In the above formula, \( F_i \) represents the receptive field of the convolution kernel in the \( i \)-th convolution layer. \( k_j \) is the size of the convolution kernel. \( r_j \) is the void rate of the \( i \)-th convolution layer. \( s_n \) is the convolution step size. Different receptive fields mean that the convolution kernel can extract features of different scales in the feature graph, connect the features output by different convolution kernels, and realize the fusion of features through \( 1 \times 1 \) convolution. In order to prevent model overfitting, each deep void convolution layer adopts a random discarding strategy. In order to reduce model parameters and avoid overfitting, the Dropout layer randomly sets some model parameters to zero according to the discard rate.

2.4. Multi-Attention

CNN can capture local features very well, but it can't capture the global connection of text. Multi-head self-attention mechanism can capture long-distance dependence well, reduce the loss of key feature information, and emphasize the global information of text sequence. The purpose of the multi-head self-attention mechanism is to capture the key information of the sequence from many aspects. The specific principle of this mechanism is to map the query matrix (Q), key matrix (K) and value matrix (V) into several different subspaces, and the subspaces are calculated separately without interfering with each other. Finally, the output of each subspace is splice together.

2.5. Focal Loss

Focal Loss can have an inhibiting effect on samples that are easy to classify, so that losses are concentrated on a small number of samples that are difficult to classify. Therefore, the network model focuses on the difficult to classify samples and increases the detection ability of the difficult to classify samples, thus increasing the overall detection accuracy. In order to reduce the influence of simple
samples on the model and increase the influence of difficult samples, the influence factor was introduced into the traditional cross entropy loss and the weight coefficient of balanced difficult samples was increased. The formula for Focal Loss:

\[
Focal \ Loss = \begin{cases} 
-\alpha(1-y')^\gamma \log y', & y = 1 \\
-(1-\alpha)y''^\gamma \log(1-y'), & y = 0 
\end{cases}
\tag{5}
\]

3. Results & Discussion

3.1. Experimental Environment
This article builds an experimental environment based on a server: the operating system Ubuntu 16.04 LTS, the internal memory is 64GB, the processor is Intel Xeon E5-2650 V4, and the CPU is 2.2GHz×8. This experiment use Keras with Tensorflow as the back end to build a convolutional neural network, use Python 3.6 as the programming language, and the experimental development tool is PyCharm 3.6.

3.2. Dataset
The experiment used two publicly available data sets available on the Internet. The MEDLINE dataset collects biomedical articles with article titles and abstracts. The data set contains a training set of 94936 articles and a test set of 48906 articles. The original data set contains more than 20,000 categories. Ten indistinguishable categories (e.g., neurology, gastroenterology, and oncology) were selected for the task of categorizing documents. The Ohsumed dataset is a subset of clinical papers from the MEDLINE database, which contains 23 categories of cardiovascular disease. The Ohsumed dataset has 13,929 documents, with varying numbers of documents per class and very irregular distribution. If a table is divided into parts these should be labelled (a), (b), (c) etc but there should only be one caption for the whole table, not separate ones for each part.

3.3. Performance Evaluation
The multi-label classification problem is similar to the binary classification problem, which can be used for performance evaluation through confusion matrix. According to the real category and predicted category, it can be divided into true category (TP), true negative category (TN), false positive category (FP) and false negative category (FN). Three indicators of accuracy, recall and F1 are used to test classification accuracy.

3.4. Results
In order to evaluate the performance of the improved CNN method in text classification, this paper introduces the traditional CNN-Softmax method, and carries out comparative experiments and analyses. Comparison results of the two algorithms are shown in Table 1. The classification accuracy of the improved CNN method proposed in this paper is better than that of CNN-Softmax method in general. The improved CNN-SVM method can obtain the deep multi-scale features, so the text classification can be better. Meanwhile, the improved CNN-SVM method can alleviate the impact of data imbalance on the classification effect and improve the overall classification effect to a certain extent.

| Model        | Classification accuracy | Kappa coefficient |
|--------------|-------------------------|-------------------|
| CNN-Softmax  | 86.14%                  | 88.7%             |
| Improved CNN | 90.88%                  | 90.56%            |

4. Conclusions
For traditional CNN, there is a large amount of computation, the problem of neglecting global connection and the problem of data imbalance in biomedical texts. In this paper, the improved
convolutional neural network and focusing loss function are used to solve the above problems, and good results are obtained. Next, the text with similar characteristics will be classified to further improve the accuracy and efficiency of classification.

References

[1] Ghiassi M., Olschimke M., Moon B. (2012) Automated text classification using a dynamic artificial neural network model. J. Expert Systems with Applications., 39 (12): 10967-10976.

[2] Yoon K. (2014) Convolutional Neural Networks for Sentence Classification. J. Eprint Arxiv.

[3] Ma M., Huang L., Xiang B. (2015) Dependency-based Convolutional Neural Networks for Sentence Embedding. In: The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing. pp. 174-179.

[4] Rie J., Tong Zh. (2015) Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. J. Eprint Arxiv.

[5] Siwei L., Liheng X., Kang L., Jun Z. (2014) Recurrent Convolutional Neural Networks for Text Classification. In: The Twenty-Ninth AAAI Conference on Artificial Intelligence. Quebec. pp.2267-2273.

[6] Ji Y. L., Franck D. (2016) Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. In: NAACL2016. San Diego. pp. 515-520.

[7] Zhou C., Sun C., Liu Z. (2015) A C-LSTM Neural Network for Text Classification. J. Computer Science., 1(4):39-44.

[8] Cao Y., Xu R., Chen T. (2015) Combining Convolutional Neural Network and Support Vector Machine for Sentiment Classification. J.

[9] Fang R., Schindelman G., Auken K. V. (2012) Automatic categorization of diverse experimental information in the bioscience literature. J. BMC bioinformatics., 13(1): 16.

[10] Cohen K. B., Hunter L. E. (2013) Text Mining for Translational Bioinformatics. J.. PLoS computational biology., 9(4): e1003044.