Longer Text Classification Based on BiLSTM-CNN Model

Dongdong Song, Baoshan Sun

Abstract— **A single CNN network has many drawbacks when processing text classification tasks. For example, when classifying longer texts, it is easy to ignore the contextual semantic information, resulting in low classification accuracy. This paper proposes a model that combines CNN and BiLSTM: BiLSTM-CNN's long text classification method. First, use the Skip-gram model in Word2Vec to map the words in the data and convert them into fixed-dimensional vectors; then use BiLSTM to capture bidirectional semantic information; finally, the features extracted by the BiLSTM model and the word embedding features are spliced as a CNN The input of, uses the convolution kernel of size 2, 3, 4 for convolution. Experiments were conducted on two public data sets of THUCNews and SougouCS. The experimental results show that the fused BiLSTM-CNN model is better than the BiLSTM and CNN models in the classification of longer texts.**

Index Terms— **CNN; BiLSTM; Skip-gram**

I. INTRODUCTION

With the development of the Internet and electronic products, people have more and more sources of text information. Faced with the ever-explosive growth of long text data, it has caused the phenomenon of excess information and lack of knowledge. So, how to efficiently manage massive amounts of disorganized data and quickly filter out valuable text information? This highlights the importance of text classification technology.

Text classification refers to learning potential rules for large-scale classification of sample data through a specific learning mechanism, and then assigning new samples to one or more categories according to the rules. The main process includes data preprocessing, text representation, feature extraction and classifier construction. Traditional text classification usually combines Bag-of-Word methods with machine learning algorithms. The bag-of-words method regards each document as composed of multiple words, which are independent of each other. Its grammar, word order and semantic information [1], but the text classification based on the bag-of-words method has problems such as high feature dimension and sparse data, which cannot accurately represent the contextual semantic information. The commonly used machine learning algorithms for classifier construction in text classification include: Support Vector Machine (SVM) [2], K-nearest neighbor (KNN) [3] and Naive Bayes (NB) [4] and other classification algorithms.

With the rapid development of today's society, the era of big data is advancing steadily, and its deep learning has achieved excellent results in image processing, speech recognition and other complex objects. Many researchers have applied deep learning to natural language processing. In the face of massive text data, in 2013 Google proposed the Word2Vec word vector tool, which can map high-dimensional word vectors to a fixed-dimensional space. Jin et al. [5] In 2014, the word vector was trained through Word2Vec, and the word embedding method was used to convert the words in the text into a fixed-dimensional word vector matrix, which was then used as the input of the convolutional neural network, and finally convolution kernels of different sizes were used for local feature extraction. , Which effectively proves the validity of the word vector. In the same year, Kalchbrenner et al. [6] K-MaxPooling pooling is designed based on the principle of MaxPooling, that is, a sliding window of a certain size is set, and the first K feature values are extracted during each sliding process. This method is gradually applied in various fields. Zhou et al. [7] In 2015, the semantic information of the context was considered, which made up for the lack of context information in CNN, combined with the advantages of CNN and LSTM, and applied it to text sentiment analysis. This research has achieved good results.

In summary, the BiLSTM-CNN model for longer text classification proposed in this paper mainly completes the following tasks: First, use the Skip-gram model in Word2Vec to map the words in the data and convert them into fixed-dimensional vectors; Then BiLSTM is used to capture bidirectional semantic information; finally, the features extracted by the BiLSTM model and the word embedding features are spliced as the input of CNN, and the convolution kernels of size 2, 3, and 4 are used for convolution. The results show that the fused BiLSTM-CNN model outperforms the BiLSTM and CNN models in news text classification.

II. BiLSTM-CNN MODEL

The BiLSTM-CNN model proposed in the article mainly includes word embedding layer, BiLSTM layer, and CNN layer. The specific network structure is shown in Figure 1.

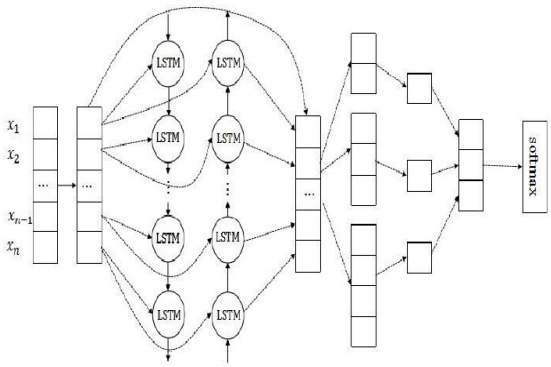

Figure 1 Flow chart of text classification based on BiLSTM-CNN model

Manuscript received September 07, 2021

Dongdong Song, School of computer science and technology, Tiangong university, Tianjin ,China

Baoshan Sun, School of computer science and technology, Tiangong university, Tianjin ,China

2.1 Word Embedding Layer

In natural language processing, in order for the computer to recognize the content of the text, the text needs to be digitally represented. In the early days, One-Hot Encoding was often used to digitally represent words. The word vector represented by this method is not only easy to cause dimensionality disasters and data sparse problems, but also cannot represent the semantic relevance between words [8].

In order to solve the problems caused by one-hot encoding, this article uses the Word2Vec model proposed by Mikolov et al. Not only can the words in the text data be converted into fixed-dimensional word vectors, but also the semantic relevance between words can be expressed. The Word2Vec model includes two word vector training models, CBOW and Skip-gram. Each model consists of an input layer, a mapping layer and an output layer. CBOW predicts the current word by surrounding words of a certain length, while Skip-gram uses the current word to predict surrounding words of a certain length[9]. In order to better represent the semantic information in the text data [10], this paper uses the Skip-gram model to map words to a fixed-dimensional space.

The specific model is shown in Figure 2.

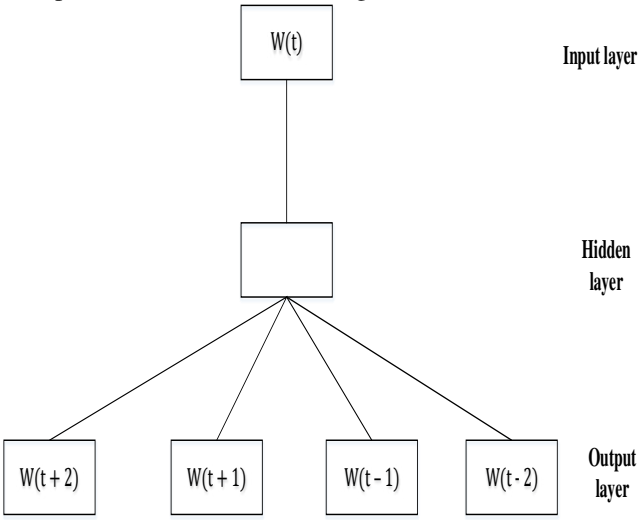

Figure 2 Skip-gram structure

2.2 BiLSTM layer

BiLSTM is a combination of forward LSTM and backward LSTM. Although the LSTM model can capture long-distance dependencies, it cannot encode before and after information [11-12]. In order to capture the before and after information, BiLSTM is used after the word embedding layer to capture richer feature information and improve the classification accuracy. The LSTM model is shown in Figure 3

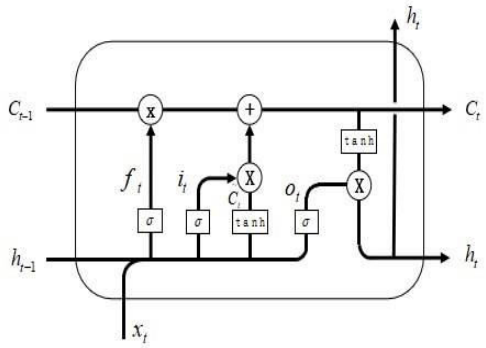

Figure 3 LSTM model structure

2.3 CNN layer

In order to further extract richer features, the features extracted by the BiLSTM layer are spliced with the features of the embedding layer, and then CNN is used for local feature extraction. Use convolution kernels of size 2, 3, and 4 for convolution, and then use K-MaxPooling to pool the features obtained after convolution fusion, where K=2, and finally concatenate the pooled features, Softmax function Used to complete text classification tasks. The CNN structure is shown in Figure 4

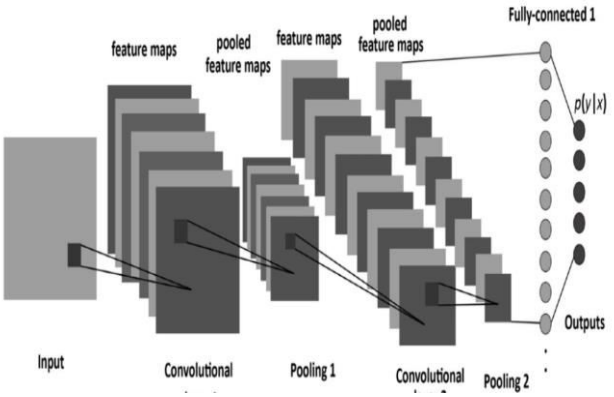

Figure 4 CNN structure

III. EXPERIMENTAL SETUP

3.1 Experiment environment

This article is based on windows10, 16GB memory, 1T hard disk, Intel Core i7 laptop for experiments. The language used in the entire experiment is Python 3.6, and the underlying framework for building the deep learning model is pytorch.

3.2 Experimental data

The experimental data used the news data THUCNews and SougouCS published by the Natural Language Laboratory of Tsinghua University and Sogou Laboratory for experiments. The THUCNews data set has a total of 740,000 documents in 14 categories, namely education, technology, fashion, finance, society, current affairs, home furnishing, constellation, games, entertainment, real estate, lottery, real estate, and sports. The number of documents in each category is not distributed. balanced. There are 17,910 documents in the SououCS data set, 9 categories, namely IT, recruitment, education, health, finance, military, culture, sports, and tourism. Each category has a total of 1990 documents. Randomly extract 70% of the training set and 30% of the test set from each label in the two data sets

3.3Experimental evaluation standards

The experiments in this article are carried out on three models, namely CNN and BiL-LSTM, BiLSTM-CNN, the model's classification effect on news text data is evaluated by three indicators: macro accuracy, macro recall, and macro F1. The macro accuracy rate represents the arithmetic average of the accuracy rates of each category, as shown in equation (1); the macro recall rate represents the arithmetic average of the recall rates of each category, as shown in equation (2); the macro F1 is each category F1 The arithmetic mean of, as shown in formula (3)

$$macro - P = \frac{1}{n}\sum_{i=1}^{n} P_i \quad (1)$$

$$macro - R = \frac{1}{n}\sum_{i=1}^{n} R_i \quad (2)$$

$$macro - F_1 = \frac{2 \times macro-P \times macro-R}{macro-P + macro-R} \quad (3)$$

3.4Experimental results and analysis

In order to better verify the effectiveness of the BiLSTM-CNN model, this paper conducts experiments on the THUCNews and SougouCS datasets, and compares them with the BiLSTM and CNN models, and passes the macro accuracy rate, macro recall rate and macro F1. The experimental results are shown in Table 1 and Table 2.

Table1 THUCNews data set experimental comparison results

Model	Maroc-P(%)	Maroc-R(%)	Maroc-F1(%)
CNN	95.48	95.50	95.49
BiLSTM	95.33	95.19	95.26
BiLSTM-CNN	96.01	95.99	96.00

Table2 Sougou CS data set experimental comparison results

Model	Maroc-P(%)	Maroc-R(%)	Maroc-F1(%)
CNN	91.61	91.60	91.60
BiLSTM	91.20	91.01	91.10
BiLSTM-CNN	92.11	92.11	92.11

It can be seen from Table 1 and Table 2 that in the THUCNews data set, BiLSTM-CNN Compared with CNN and BiLSTM models, the macro accuracy rate is increased by 0.53% and 0.68%, and the macro recall rate is increased by 0.49% and 0.80% and macro F1 respectively. respectively. Increased by 0.51% and 0.74% respectively. On the SogouCS data set, compared with the CNN and BiLSTM models, BiLSTM-CNN has improved macro accuracy by 0.50% and 0.91%, macro recall by 0.51% and 1.10%, and macro F1 by 0.51% and 1.01%. The improvement effect on the SogouCS dataset is significantly higher than the THUCNews dataset, mainly because the number of samples in each category in the THUCNews dataset is not balanced, and secondly because of the hierarchical structure between certain categories, such as sports categories including football categories . Based on the above experimental results, it can be concluded that the BiLSTM-CNN model is better than the single CNN and BiLSTM models in news text classification.

CONCLUSION

The BiLSTM-CNN model for news text classification proposed in this paper uses the Word2Vec word vector model to map words to a fixed-dimensional space, which solves the problem of high feature dimension and semantic irrelevance brought by traditional One-hot encoding. In order to better integrate with temporal features, first use BiLSTM to learn temporal features from the embedding layer; then extract deep features through convolution kernels of different sizes; and finally fuse the learned features. Experiments show that the BiL-STM-CNN model proposed in this paper is superior to the single CNN and BiLSTM models in terms of news text classification. In the follow-up work, the news text with hierarchical structure categories will be studied to improve the accuracy of classification.

REFERENCES

[1] Zhang Y , Kong W , Dong Z Y , et al. Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network[J]. IEEE Transactions on Smart Grid, 2019.

[2] Wang, Xuan, Xu, et al. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN[J]. Expert Systems with Application, 2017.

[3] Lin B Y , Xu F , Luo Z , et al. Multi-channel BiLSTM-CRF Model for Emerging Named Entity Recognition in Social Media[C]// Proceedings of the 3rd Workshop on Noisy User-generated Text. 2017.

[4] Luo L , Yang Z , Yang P , et al. An Attention-based BiLSTM-CRF Approach to Document-level Chemical Named Entity Recognition[J]. Bioinformatics, 2017(8):8.

[5] Lin C Y , Xue N , Zhao D , et al. A Convolution BiLSTM Neural Network Model for Chinese Event Extraction[C]// Springer International Publishing. Springer International Publishing, 2016:275-287.

[6] Shin H C , Roth H R , Gao M , et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning[J]. IEEE Transactions on Medical Imaging, 2016, 35(5):1285-1298.

[7] Perez-Munuzuri V , Perez-Villar V , Chua L O . Autowaves for image processing on a two-dimensional CNN array of excitable nonlinear circuits: flat and wrinkled labyrinths[J]. IEEE Trans.circuits Syst.i Fundam.theory Appl, 1993, 40(3):174-181.

[8] Yiming, Yang. An Evaluation of Statistical Approaches to Text Categorization[J]. Information Retrieval, 1999.

[9] [9] C Apté, Damerau F , Weiss S M . Automated learning of decision rules for text categorization[J]. Acm Transactions on Information Systems, 1994, 12(3):233-251.

[10] A Comparative Study on Feature Selection in Text Categorization[C]// Proc Int Conference on Machine Learning. 1997.

[11] Context-sensitive learning methods for text categorization[J]. Acm Transactions on Information Systems, 1999, 17(2):141-173.

[12] Lewis D D , Info C F , Lang S , et al. A Comparison of Two Learning Algorithms for Text Categorization[J]. third annual symposium on document analysis & information retrieval, 1996.