The Automatic Text Classification Method Based on BERT and Feature Union

Wenting Li, Shangbing Gao * , Hong Zhou, Zihe Huang, Kewen Zhang and Wei Li
Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian 223003, P.R.China
Corresponding author: luxiaofen_2002@126.com

***Abstract*—For the traditional model based on the deep learning method most used CNN(convolutional neural networks) or RNN(Recurrent neural Network) model and is based on the dynamic character-level embedding or word-level embedding as input, so there is a problem that the text feature extraction is not comprehensive. In the development environment of the Internet of Things, A method of Automatic text classification based on BERT(Bidirectional Encoder Representations from Transformers) and Feature Fusion was proposed in this paper. Firstly, the text-to-dynamic character-level embedding is transformed by the BERT model, and the BiLSTM(Bi-directional Long-Short Term Memory) and CNN output features are combined and merged to make full use of CNN to extract the advantages of local features and to use BiLSTM to have the advantage of memory to link the extracted context features to better represent the text, so as to improve the accuracy of text classification task. A comparative study with state-of-the-art approaches manifests the proposed method outperforms the state-of-the-art methods in accuracy. It can effectively improve the accuracy of tag prediction for text data with sequence features and obvious local features.**

***Index Terms*—NLP; BERT; BiLSTM; CNN; Feature Union; Text classification**

I. INTRODUCTION

With the sharp increase of online text information data, text classification plays a vital role in information. It is a key technology to process large scale text information and drives the direction of information in the direction of automation.There are many researches on related implementation methods. A graph-CNN based deepth learning model is proposed by Peng H et al [1].It's main idea is to convert texts to graph-of-words firstly, and then uses graph convolution operations to convolve the word graph.Graph-of-words representation of texts has the advantage of capturing non-consecutive and long-distance semantics. CNN models have the preponderance of learning different level of semantics. Cross-domain sentiment classification via a bifurcated-LSTM is proposed by Ji J et al [2]. A bifurcation LSTM is designed by using attention-based LSTM classifier with enhanced data set and orthogonal constraints. Domain-invariant sentiment features can be extract from the source domain by the Bifurcated-LSTM to perform sentiment analysis in different target domains and the system leads to significant performance improvement.

In the past text classification tasks, the traditional learning method is based on the dynamic character-level embedding or word-level embedding as the downstream task inputs, the dynamic character-level embedding or word-level embedding lacks the semantic information of the context. And most of the traditional method of deep learning based on deep learning is based on CNN or RNN. The former is difficult to extract the serialization features of text, while the latter has poor encoding ability for words.

In order to solve the above problems existing in the prior art, *the automatic text classification method based on BERT and Feature Union* is provided by this paper. Overall, our main contributions are two folded: (1) The traditional deep learning method is based on non-dynamic word vector or word vector as input. The word vector or word vector cannot be changed according to its context. The information coverage is relatively simple. This paper uses the BERT pre-training model to generate dynamic word vectors with richer contextual semantic information. (2) In this paper, the advantages of CNN and BiLSTM are complemented by feature fusion method, and the serialization features of text can be extracted, and the coding ability of words is better.

The paper is organized as follows. In section 2, we describe the main components of the improved model in detail. Section 3 presents the experimental comparison between the model of this paper and other models. The paper closes in section 4 with a systematic conclusion for this paper and a discussion for future work.

II. METHOD

Text categorization is one of the main tasks of NLP (Natural Language Processing) and plays a significant role in information processing.

The model is proposed creatively in this paper. Each charactor in the text is vectorized by the BERT pre-training language model . The obtained charactor vector sequence is deeply coded twice that the charactor vector sequence are input into the CNN [4] and BiLSTM models respectively, the two features of the output are merged, and then the final prediction label is output through the fully connected layer and softmax. Detailed flow chart is shown in the Fig. 1.

A. Pre-training Language Model

BERT [5] is the state of the art model released by Google in October 2018. Compared to the previous pre-training model, it can consider both sides of a character at the same time. It can be seen that the BERT implements the semantic information of the context in the true sense.

Each token in BERT is represented by the addition of the three parts:

978-1-7281-2583-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICPADS47876.2019.00114

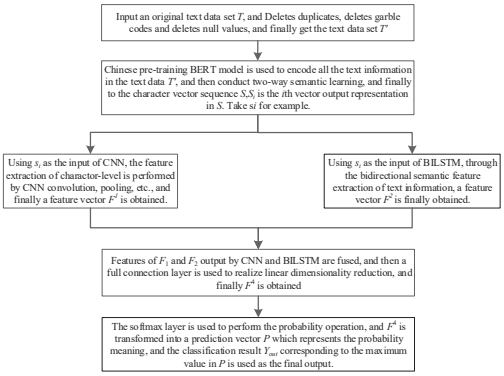

Fig. 1. The overall frame diagram of the improved model

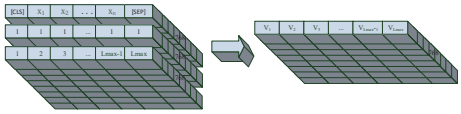

Fig. 2. The input representation of BERT model

- Token embeddings represents a vector representation that converts each charactor into a fixed dimension.
- Segment embeddings is used to split the text information, indicating different samples, here is a single input.
- Position embeddings represents location information

Each token of each text message in the input text data set T' is sent to the Token Embedding layer, the Segment Embeddings layer, and the Position Embeddings layer of the BERT. The vector code V_1, the segment code V_2 and the position code V_3 are respectively obtained, and the obtained V_1, V_2 and V_3 are added and input into the BERT. The input representation of BERT model is shown in Fig. 2.

B. Downstream Network Structure

The convolutional neural network CNN [7] in this paper. First, the similarity with some keywords is calculated in the process of convolution, and then through the max-pooling layer to find out whether the keywords that get attention of the model appear in the entire input text, and the similarity between the most similar keywords and the convolution kernel. The convolutional neural network structure is shown in Fig. 3.

$k = 768$ is the dimension of the character vector and $Lmax = 200$ is the maximum length of the input.

The input of convolutional neural network CNN is $V(W_i)$. In the first convolutional layer, $100 * 5$ convolution with a quantity of 128 are used to convolute the input V_i to obtain the feature f_1;

$$f_i = W \cdot V_i + b \tag{1}$$

Inputting the obtained f_1 into the activation function Relu() to obtain the feature f_2. The calculation formula is as follows:

$$f_2 = \max(0, f_1) \tag{2}$$

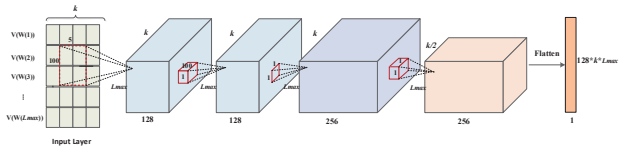

Fig. 3. Convolutional neural network structure diagram

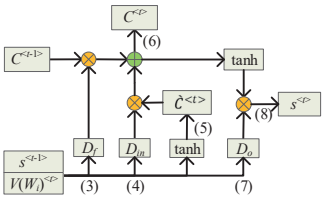

Fig. 4. BERT model input representation

The feature f_2 obtained in the previous step is inputted to the max-pooling layer, and use the 1×101 pooling kernel to perform the pooling with the step length of 1, so as to obtain the output feature f_3.

Repeating the above steps and replacing some of the parameters: The convolution layer is convolved with 1×1 convolution kernel of the quantity of 256 , and the max-pooling layer is pooled with 1×1 pooling kernel with step size of 2, so as to obtain the output f_4. Finally, the Flatten layer is used to compress the multidimensional feature f_4 into a one-dimensional vector, so as to obtain the final output feature *F1* of CNN.

Although CNN has a strong superiority in the extraction of local features, but it can not know which keywords appear several times and in which order.

Context sensitive information can be obtained by standard RNNs. However, the range of context information is very limited which it accesses and it is easy to cause the gradient explosion live gradient disappears. Therefore, the LSTM (Long and Short Memory Network) structure was born. The structure diagram of the LSTM unit is shown in Fig. 4:

The following formula can calculate the value of each neuron in the Fig. 4:

$$D_f = sigmoid(W_f[V(W_i), s^{t-1}] + b_f) \tag{3}$$

$$D_{in} = sigmoid(W_{in}[V(W_i), s^{t-1}] + b_{in}) \tag{4}$$

$$\grave{C}^{<t>} = tanh(W_c[V(W_i), s^{t-1}] + b_c) \tag{5}$$

$$C^{<t>} = D_f * C^{<t-1>} + D_{in} * \grave{C}^{<t>} \tag{6}$$

$$D_o = sigmoid(W_o[V(W_i), s^{t-1}] + b_o) \tag{7}$$

$$s^{<t>} = D_o * tanh(C^{<t>}) \tag{8}$$

In the above formulas, $C^{<t-1>}$ is the cell state information of the previous moment. W_f, W_{in}, b_f, and b_{in} are the weight matrix and bias term of the forgetting gate and the input gate respectively. The output D_f of the forgetting gate and the output D_{in} of the input gate can be obtained by the formulas

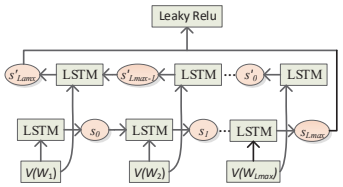

Fig. 5. BiLSTM network structure diagram

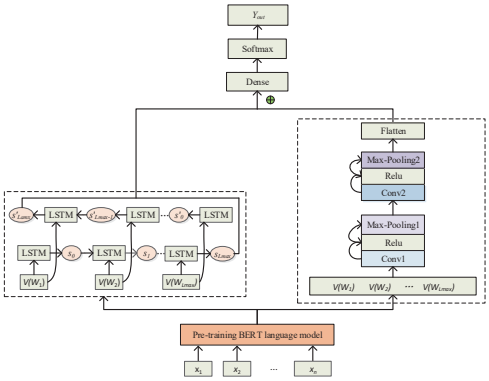

Fig. 6. The improved network structure diagram

(3) and (4) respectively. The unit state $\grave{C}^{<t>}$ input at time t can be obtained by formula (5), W_c and b_c are the weight matrix and the bias term of the input unit at time t respectively, and the updated cell state information $C^{<t>}$ can be obtained by formula (6). The output D_o of the output gate can be obtained by the formula (7), and finally the hidden layer state $s^{<t>}$ at time t can be calculated by the formula (8).

Although the problems of RNN can be solved perfectly by LSTM, the semantic information between the contexts cannot be grasped by the LSTM. BiLSTM [8] is a variant of the Long Short Term Memory Network (LSTM). The problem of context correlation in LSTM is fully solved. The BiLSTM network structure is shown in Fig. 5.

The advantages of CNN in extracting local features and the strength of BiLSTM [9] with memory are fully utilized in the feature fusion of downstream tasks in this paper. The improved model of this paper is shown in Fig. 6.

The character vector sequence S is deeply coded twice. In the CNN model, the input character vector sequence is operated by two convolutions, two activations, and two max-pooling methods, and then the obtained multidimensional features are transformed into one-dimensional features by the Flatten layer, and finally a feature vector F^1 is obtained. $F^1 = \{F_1^1, F_2^1, \ldots, F_k^1\}$, $k = 98304$ is the number of neurons output by the second pooling layer of CNN.In the BiLSTM model, the input word vector sequence is subjected to a bidirectional operation on the sequence, and the obtained output is subjected to the operation of the nonlinear activation layer as another feature vector F^2, $F^2 = \{F_1^2, F_2^2, \ldots, F_m^2\}$, $m = 768$ is the number of BiLSTM hidden layer neurons.

The two eigenvectors F^1 and F^2 are fused. Firstly, the two eigenvectors are spliced to obtain a new eigenvectors, $F^3 = \{F_1^3, F_2^3, \ldots, F_l^3\}$, $l = k + m$. Then, the eigenvector F^3 passes through the full connection layer to realize linear dimensionality reduction, and the dimensionality is finally reduced to 10 (corresponding to 10 prediction results), so as to obtain eigenvector $F^4 = \{F_1^4, F_2^4, \ldots, F_{10}^4\}$.

Finally, the softmax layer is used to perform the probability operation. The obtained eigenvector F^4 as input and input to the softmax layer, so that each real number in the input vector is mapped to a real number between 0 and 1. The obtained output is the probability prediction vector $P = \{p_1, p_2, \ldots, p_f, \ldots, p_{10}\}$, $f \in [1, 10]$, $p_f = e^{F_f^4}/(\sum_{f=1}^{10} e^{F_f^4})$ represents this text is the probability of class f.

TABLE I
PARAMETER SETTING OF BERT MODEL

Parameter Name	*Parameter Values*	*Parameter Name*	*Parameter Values*
hidden_size	768	learning_rate	5e-5
Lmax	200	keep_prob	0.9
optimizer	Adam	loss_fn	softmax

TABLE II
PARAMETER SETTING OF CNN MODEL

Parameter Name	*Parameter Values*	*Parameter Name*	*Parameter Values*
Conv1	100*5	Conv2	1*1
Conv1_channel	128	Conv2_channel	256
keep_prob	0.5	active_fn	relu

III. EXPERIMENTS

A. Experimental environment

The experimental environment in this paper is an AI server configured with 2× Intel Xeon 6148,512g memory, 4×1.9 t SSD hard disk, raid card, 2× ten thousand network card, 8×Tesla V100 card, 2× double port 100Gbps HCA card, 3000W $1 + 1$ redundant server power. The experimental data in this paper is derived from the recruitment information text data of the future worry-free 51job website during the period of $2018 - 2019$. The total amount of data in the experiment is 3541311.

B. Experimental parameters

The experimental text dataset of this paper is divided into training set, verification set and test set according to the ratio of 6:2:2.The parameter settings in the BERT model are shown in TABLE I. The parameters of the convolutional neural network CNN and BILSTM are shown in TABLE II and TABLE III.

C. Experimental results

In the experimental compaison of CNN, BILSTM, BERT, BERT+CNN and BERT+BILSTM, the model proposed in this paper is superior to other models in all aspects.. The accuracy, recall rate and F1-Score of the text classification

TABLE III
PARAMETER SETTING OF BILSTM MODEL

Parameter Name	*Parameter Values*	*Parameter Name*	*Parameter Values*
forward_hidden _size	384	output_keep_prob (if training)	0.5 (1.0)
backward_hidd	384	active_fn	leaky_relu
input_keep_pro	1.0	Alpha(leaky_relu)	0.2

TABLE IV
STATISTICAL TABLE OF EXPERIMENTAL RESULTS

Algorithm	*Precision*	*Recall*	*F1-score*
CNN	0.74	0.66	0.69
BiLSTM	0.84	0.75	0.79
BERT	0.92	0.92	0.92
BERT+CNN	0.93	0.93	0.93
BERT+BILSTM	0.92	0.92	0.92
BERT+(BILSTM-CNN)	0.96	0.96	0.96

of 10 labels reached the optimal 96% classification effect. The experimental comparison results are shown in TABLE IV.

In order to more intuitively see the effect of the improved model, a confusion matrix heat map is generated here using a test data set with a quantity of 70801.

As shown in Fig. 7, the corresponding probability value is more and more higher from the purple-blue-green-yellow transition. From the picture,we can find the improved model is very effective in the accuracy of automatic text classification.

In the Figure 8, It can be clearly seen from the following figure that almost all the sample numbers are concentrated on the diagonal line. That is to say, most of the samples are predicted correctly.

IV. CONCLUSION

Although the improved model has been greatly improved in accuracy, the time cost problem of the model has not been effectively improved, and the network structure in the downstream task needs to be further explored. In the follow-up work study. I expect that we can continue to improve and upgrade the methods of this article, so that this method can be more effectively applied to various needs and provides people with a more convenient life.

ACKNOWLEDGEMENTS

This work is sponsored by the National Key R&D Program of China (No. 2018YFB1004904), the Huai'an science and technology project (No. HAC201705, No.HAB201803), the key project of Jiangsu Provincial Department of Education (No.18KJA520001), six talent peaks project in Jiangsu Province (XYDXXJS-011), Jiangsu 333 engineering research funding project (BRA2016454).

REFERENCES

[1] Peng H, Li J, He Y, et al. Large-scale hierarchical text classification with recursively regularized deep graph-cnn[C]//Proceedings of the 2018 World Wide Web Conference. International World Wide Web Conferences Steering Committee, 2018: 1063-1072.

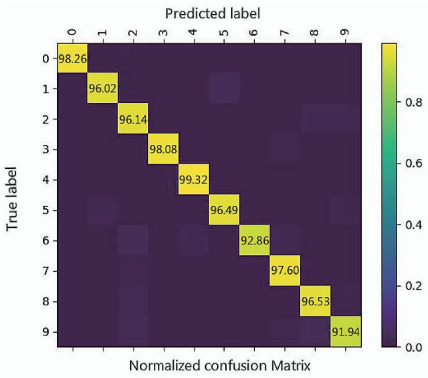

Fig. 7. Confusion matrix heat map based on probability

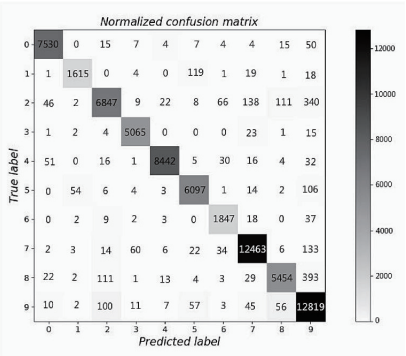

Fig. 8. Confusion matrix heat map based on quantity

[2] Ji J, Luo C, Chen X, et al. Cross-domain sentiment classification via a bifurcated-LSTM[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2018: 681-693.

[3] Xiulong Liu, Xin Xie, Shangguang Wang, Jia Liu, Didi Yao, Jiannong Cao, Keqiu Li, Efficient Range Queries for Large-scale Sensor-augmented RFID Systems, IEEE/ACM Transactions on Networking (TON), in press, 2019.

[4] Hassan A, Mahmood A. Convolutional recurrent deep learning model for sentence classification[J]. Ieee Access, 2018, 6: 13949-13957.

[5] 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.

[6] Xiulong Liu, Jiannong Cao, Yanni Yang, Wenyu Qu, Xibin Zhao, Keqiu Li, Didi Yao, Fast RFID Sensory Data Collection: Trade-off Between Computation and Communication Costs, IEEE/ACM Transactions on Networking (TON), 27(3), 2019.

[7] Yoon H J, Robinson S, Christian J B, et al. Filter pruning of convolutional neural networks for text classification: a case study of cancer pathology report comprehension[C]//2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). IEEE, 2018: 345-348.

[8] Ghaeini R, Hasan S A, Datla V, et al. Dr-bilstm: Dependent reading bidirectional lstm for natural language inference[J]. arXiv preprint arXiv:1802.05577, 2018.

[9] Li C, Zhan G, Li Z. News Text Classification Based on Improved Bi-LSTM-CNN[C]//2018 9th International Conference on Information Technology in Medicine and Education (ITME). IEEE, 2018: 890-893.