2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA) | 978-1-6654-2490-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CAIBDA53561.2021.00036

News text classification based on Bidirectional Encoder Representation from Transformers

Lin Deping
School of Information Engineering
Beijing Institute of Graphic Communication
Beijing, China
1316147709@qq.com

Wang Hongjuan*
School of New Media
Beijing Institute of Graphic Communication
Beijing, China
229359989@qq.com

Liu Mengyang
School of Information Engineering
Beijing Institute of Graphic Communication
Beijing, China
1367274860@qq.com

Li Pei
School of Information Engineering
Beijing Institute of Graphic Communication
Beijing, China
1365588043@qq.com

***Abstract*—In order to accurately and efficiently obtain information useful to us, people are paying more and more attention to the problem of data redundancy caused by excessive data information. In recent years, domestic and foreign researchers have proposed various frameworks for different natural language processing tasks, and different frameworks have different advantages and disadvantages. One of the classic problems in the field of natural language processing is text classification. News text classification is an important task that is easy to attract everyone's attention in our daily lives. This experiment is based on the BERT model under the Transformer framework to classify the news text data set. The same news text data set is compared with the RNN's long and short-term memory network. The evaluation index uses the general accuracy and loss value of the model classification. Experimental results show that the classification accuracy of the BERT model is significantly higher than that of the long and short-term memory network.**

Keywords-component; news text classification; LSTM; Bert

I. INTRODUCTION

Natural language processing is an important research direction in the current artificial intelligence field, and it has an extremely important position. Its purpose is to study how to convey human language or human ideas to computers, so as to realize the communication between humans and computers. [1]. The transmission of human information requires the use of a large number of languages, and all kinds of human inventions, innovative thinking, and experience knowledge need to be passed on in the form of languages. Therefore, the core of artificial intelligence is natural language processing, which can be applied to many aspects of human real life, such as machine translation, public opinion monitoring, automatic summarization, opinion extraction, text classification, etc. [2].

And what we are going to study today is one of the important aspects: the news text data set is classified based on the BERT model under the Transformer framework, and the same news text data set is compared with the RNN's long and short-term memory network. Through experiments, we obtain evaluation indicators such as the accuracy of label classification of news text data sets, and then compare the two experimental results to compare the advantages and disadvantages of the two methods, so as to screen out more efficient and accurate classification methods for human.

II. RELATED THEORIES AND TECHNOLOGIES

A. Recurrent neural network

Recurrent neural network is a kind of recurrent neural network, which takes sequence data as input, recursively in the evolution direction of the sequence, and connects all nodes in a chain (recurrent unit)[3]. The input of the convolutional network is only the input data X, and in addition to the input data X, the output of each step of the cyclic neural network will be used as the input of the next step, and so on, and each time the same activation function and parameters are used. A typical recurrent neural network structure model is shown in Figure 1. The right side is the network structure diagram after expanding it in time series.

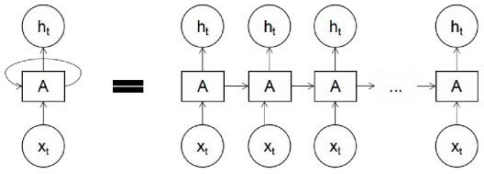

Figure 1. Basic recurrent neural network structure

B. Long short-term memory network

RNN is evolving in the direction of serialization, each time it is propagated through only one parameter, and when the amount of input information is large and complex, one parameter as input is far from enough. In order to solve this problem, long-term and short-term Memory network (Long Short Term Memory, LSTM). LSTM is a gating algorithm, and the LSTM unit contains 3 gates: input gate, forget gate and output gate [4]. This mechanism can selectively input and output required information, or selectively forget unnecessary

978-1-6654-2490-5/21/$31.00 ©2021 IEEE
DOI 10.1109/CAIBDA53561.2021.00036

information. LSTM is an improved structure of RNN. Compared with the recursive operation of the hidden layer of the recurrent neural network, the node structure of the hidden layer of the long short-term memory network is composed of three gated units and a self-loop is established inside the LSTM unit. The structure of the LSTM network unit is shown in Figure 2.2 below:

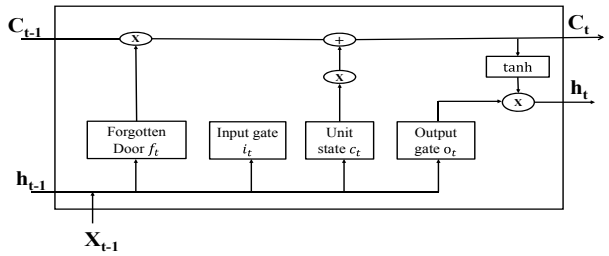

Figure 2. Long short-term memory network unit structure

In detail, the input gate determines the input of the current time step and the update of the system state's internal state; the forgetting gate determines the previous time step, and the internal state updates the internal state of the current time step; the output gate determines the internal state updates to the state of the system[5]. The update method of the LSTM unit is as follows:

$$h^{(t)} = g_0^{(t)} f_h\left(S^{(t)}\right) \quad (1)$$

$$S^{(t)} = g_f^{(t)} S^{(t-1)} + g_i^{(t)} f_s\left(wh^{(t-1)} + uX^{(t)} + b\right) \quad (2)$$

$$g_i^{(t)} = Sigmoid\left(w_i h^{(t-1)} + u_i X^{(t)} + b_i\right) \quad (3)$$

$$g_f^{(t)} = Sigmoid\left(w_f h^{(t-1)} + u_f X^{(t)} + b_f\right) \quad (4)$$

$$g_o^{(t)} = Sigmoid\left(w_o h^{(t-1)} + u_o X^{(t)} + b_o\right) \quad (5)$$

f_h, f_s are the activation functions of the system state and internal state, they just hyperbolic tangent functions, g is the gate control updated with time step, essentially a feedforward neural network with the Sigmoid function as the activation function, The reason why use the Sigmoid function is that its output is in the [0,1] interval, which is equivalent to a set of weights. The subscripts *i*, *f*, *o* represent input gate, forget gate and output gate.

C. *Transformer model as a whole*

The Transformer model was first proposed by Google in the article "Attention is all you need" in 2017[6]. In the paper, this model is mainly used to overcome the problem of too long traditional network training time in machine translation tasks, and it is difficult to achieve parallel computing. Later, because the method is better than traditional RNN and LSTM in the extraction of word order features, it has been gradually applied to various fields, such as the GPT model and the latest upstart Bert model produced by Google.

Compared with the traditional RNN, LSTM or attention concentration mechanism, the Transformer model has abandoned the previous time series structure, which is why it can be paralleled well and the training speed is faster. More accurately, it is actually an encoding mechanism that includes both semantic information (Multi-Head Attention) and position information (Positional Encoding)[7].

Transformer replaces the RNN structure in Seq2seq by using Attention extensively, so that the entire model can be calculated in parallel (RNN can only calculate the next time step after the t-1 time step calculation is completed). Like all Seq2seq, the entire model is also divided into two parts: Encoder and Decoder (corresponding to the left and right in the figure respectively). Figure 3 shows the overall structure of the Transformer model.

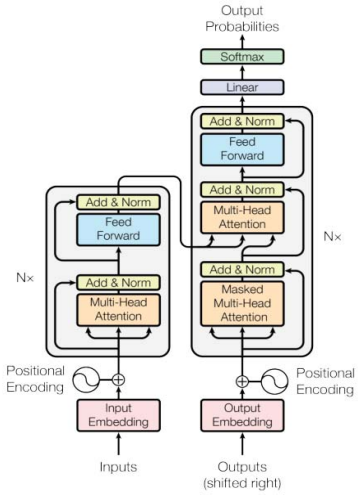

Figure 3. The overall structure of the Transformer model

D. *BERT model*

The full name of the BERT model is: Bidirectional Encoder Representations from Transformer. BERT's model architecture is based on Transformer, which implements a multi-layer bidirectional Transformer encoder. Two-way means that when processing a word, the information before and after the word can be considered, so as to obtain the semantics of the context.

The goal of the BERT model is to use large-scale unlabeled corpus training to obtain a representation of the text that contains rich semantic information, and then fine-tune the semantic representation of the text in a specific NLP task, and finally apply it to the NLP task [8]. The main input of the BERT model is the original word vector of each word in the text. This vector can be initialized randomly, or pre-trained with algorithms such as Word2Vector as the initial value. The output is a vector representation of each word in the text fused with the semantic information of the full text. Figure 4 below shows a schematic diagram of the BERT model.

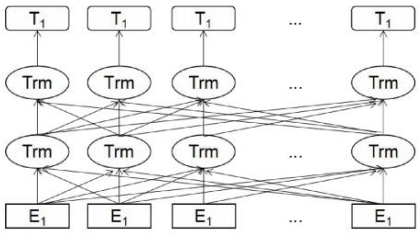

Figure 4. BERT model

In Figure 4 above, Ei refers to a single word, Ti refers to the finally calculated hidden layer, after the attention matrix and attention weighting, after this operation, every word in the sequence contains the information before the word and the information after the word, each word, after the attention mechanism and weighting, the current word is equivalent to using all the other words in this sentence to express, each word contains the information of all the components in this sentence [9].

BERT is actually a language model. Language models are usually trained using large-scale text corpus that has nothing to do with specific NLP tasks. The goal is to learn what the language itself should be like. This is like when we learn Chinese, English and other language courses, we need to learn how to choose and combine the vocabulary which we have already mastered to generate a smooth text[10]. Returning to the BERT model, the pre-training process is to gradually adjust the model parameters, so that the semantic representation of the text output by the model can describe the essence of the language, which is convenient for subsequent fine-tuning for specific NLP tasks [11].

III. COMPARISON EXPERIMENT OF NEWS TEXT CLASSIFICATION

A. News text classification experiment preparation

There is relatively little research and comparison work on the classification of news texts, so the comparison of the effects of the popular classification models has great practical significance for the classification of news texts. This article is mainly aimed at the news text data set, through the actual training experiment of the classic RNN model and the emerging BERT model to analyze the efficiency and performance of the two.

1) Lab environment

The experimental hardware environment required for the classification of news in this article is shown in the following Table I. Due to the limitation of the environment configuration of the computer used, the experimental software environment this time is provided by Google Labs as the deep learning framework Training tool. The GPU used is NVIDIA GeForce MX250 7.9GB.

TABLE I. HARDWARE CONFIGURATION TABLE

Configuration	Parameter
CPU	i7-10510U 1.8GHZ
RAM	8GB
GPU	NVIDIA GeForce MX250
Operating System	Windows 10

2) Experimental data

This experiment uses the "ag_news_subset" data set for model training. The news text data set contains more than one million news articles which is provided by the academic community for data mining (clustering, classification, etc.), information retrieval (ranking, search, etc.), xml, data compression, data streaming and any other non-commercial activities. AG's news topic classification data set was constructed by Zhang Xiang from the above data set [12]. AG's news topic classification data set is constructed by selecting the 4 largest categories from the original corpus. Each course contains 30,000 training samples and 1,900 test samples. The total number of training samples is 120,000, and the test samples are 7,600.

3) Evaluation index

The evaluation index of this experiment adopts the general accuracy rate and loss rate. Accuracy is expressed as the ratio of the number of correctly classified samples to the total number of samples on the test data set, and loss is the difference between the predicted value and the true value obtained when the model is built.

B. Experiment procedure

1) Experimental parameter settings

This article mainly uses two training models, RNN and BERT, to compare news text classification. The common parameters used in the experiments of these two models are shown in Table II below.

TABLE II. THE SAME PARAMETERS OF THE TWO MODELS

Parameter name	Parameter meaning	Parameter value
Dropout	Random dropout probability	0.1
num_classes	Number of classification categories	4
vocab_size	Vocabulary size	75376
max_epochs	The maximum number of iterations	10
batch_size	Batch size	64
learning_rate	Learning rate	1e-4
embedding	Word vector dimension	128

2) Experimental training

This experiment is conducted on the Google laboratory for network training. After configuring the tensorflow environment, download the required data set, define the RNN and BERT training models respectively, and then substitute the same news text data set for model training.

3) Comparative analysis of experimental results

This section uses different deep learning methods to train the same news text classification data set AG, and obtains the classification accuracy and loss rate of the tested sample set. The data results of the two training models are compared by drawing a line chart.

As shown in Figure 5, for the loss value, both RNN and BERT models have overfitting during the training process. However, in the first three training cycles without overfitting, the loss value of the BERT model is significantly lower than that of the RNN model.

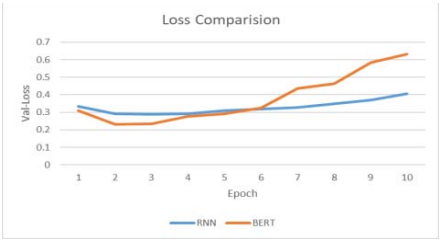

Figure 5. Comparison of loss values

As shown in Figure 6, in terms of accuracy, the verification accuracy of RNN and BERT models rose slowly during the entire training process and then occasionally fell, and eventually stabilized, but the verification accuracy of the BERT model was significantly higher than the RNN model.

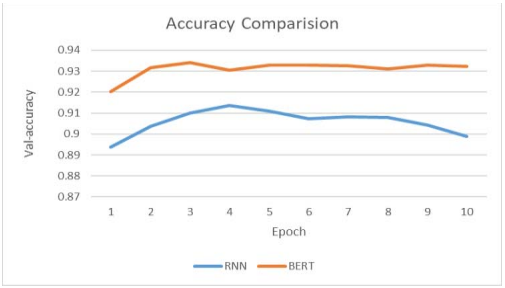

Figure 6. Accuracy comparison

By comparing the experimental results of classification training on the same news text data set by the RNN and BERT models, the classification accuracy of the BERT model is higher than the RNN model, and its classification loss value is lower than the RNN model, so the classification effect of the BERT model is better than the RNN model.

IV. Summary and Outlook

A. *Paper summary*

In this article, we first briefly introduce the theoretical knowledge and technology related to the neural network model used in this news text classification. Then through actual experimental training, the experimental environment is configured in the Google laboratory and the news text data set is used for training, and the accuracy and loss rate obtained by training the RNN and BERT model on the same news data set are obtained. The experimental results of the two are analyzed and compared, so the accuracy and advancement of the BERT model in news text classification are verified.

B. *Future outlook*

Although this paper compares the text classification model based on the news text data set, and verifies the excellence of the two models based on the comparative analysis of the experimental results, it does not propose specific and feasible improvements and optimization methods. From an algorithm perspective, there are still big shortcomings. From the experimental results, the resulting over-fitting phenomenon is also an area where experimental training needs to be improved.

In the coming days, we will conduct further research. For overfitting in training, we will reduce the training time, increase the size of the data set, cross-validate the data set, and reduce the complexity of the model through regularization in order to reduce over-fitting. For the recurrent neural network, we will start with how to improve the parallel computing capability; for the BERT model, although it is the current mainstream method and the performance is better than the traditional recurrent neural network , but the future research prospects of the BERT model are still broad. We are considering how to reduce the number of parameters without reducing the accuracy, so as to improve the work efficiency of the BERT model.

Acknowledgment

This work was supported by Beijing Natural Science Foundation (Grant No. 1212010),Beijing Municipal Party Committee Organization Department Youth Top-notch Project (Grant No. 10000200527) and the National Natural Science Foundation (Grant No. 11603004).

References

[1] Adnan Muhammad. A Methodology for Comparison of User Reviews with Rating of Android Apps using Sentiment Analysis[D]. Southwest University of Science and Technology, 2020.

[2] He Kai. Research and application of text classification based on natural language processing [D]. Nanjing University of Posts and Telecommunications, 2020.

[3] Tang Qinting. Research and implementation of network news text classification system based on deep learning [D]. Beijing University of Posts and Telecommunications, 2020.

[4] Yang Junzhuo. Improvement and application in text classification based on RNN [D]. Jilin University, 2020.

[5] Haochi Zhang，Tongyu Zhu. Aircraft Hard Landing Prediction Using LSTM Neural Network[P]. Computer Science and Intelligent Control,2018.

[6] Huang Lei, Du Changshun. Research on text classification based on recurrent neural network [J]. Journal of Beijing University of Chemical Technology (Natural Science Edition), 2017, 44(01): 98-104.

[7] Lin Wenxing. Research on text classification based on structure optimization recurrent neural network [D]. Chongqing University of Posts and Telecommunications, 2019.

[8] Xu Xu. Research on Improved Method of Recurrent Neural Network [D]. Shanghai Jiaotong University, 2017.

[9] Luo Jiahuan. Research on text summary generation technology based on Transformer model [D]. South China University of Technology, 2020.

[10] Wang Nanti. Research on improved text representation model based on BERT [D]. Southwest University, 2019.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[P]. Google AI Language,2019.

[12] Xiang Zhang, Junbo Zhao, Yann LeCun. Character-level Convolutional Networks for Text Classification[D]. New York University,2016