Text classification method based on LSTM

Qinghua Jiang[1]

[1] Youlian shipyard (Shekou) Co., Ltd
zty224@126.com

Abstract. With the rapid development of Internet technology and the explosive growth of social media, a large amount of information continues to be generated, of which the amount of text information is the largest. The main feature of various Chinese short text messages such as news headlines and instant messages is sparsity, which is only composed of a few to dozens of words, and the content of effective information packets is very small. As a result, the samples with sparse features and high feature set dimensions are difficult to provide key and accurate features for text classification learning. This paper mainly studies the application of deep learning in the field of Chinese text classification, and proposes a text classification model based on word level and character level mixed features.

Keywords: Text Classification, Deep Learning.

1 Introduce

The rapid development of Internet technology and the explosive growth of social media continue to produce text, pictures, audio, video and other information, of which the amount of text information that people can easily and conveniently obtain is also the largest. The main feature of various Chinese short text messages such as news headlines, microblogs, instant messages and online comments is sparsity, which is only composed of a few to dozens of words. For example, the length of news headlines usually does not exceed 40 characters, and the content of effective information packets is very small, The samples with sparse features and high feature set dimensions are difficult to provide key and accurate features for text classification learning. Secondly, when short text information appears on the Internet, it has the characteristics of real-time update, fast refresh speed and large number of text. These text information provide important data sources for the research in the fields of information retrieval, personalized news recommendation, relationship extraction W and user intention analysis. Therefore, how to classify text information quickly and effectively becomes very important, and text classification technology came into being. In the Internet era, people have been used to obtaining the information they need on search engines such as Google and Baidu through keywords. Many online libraries, online publishing houses and information portals provide users with index text through text databases. In the process of organizing, analyzing and managing a large number of text information in the form of electronic documents, text

EAI Urb-IoT 2021, December 20-21, Shenzhen, People's Republic of China

DOI 10.4108/eai.20-12-2021.2315016

classification technology helps users locate the required information quickly and accurately, and plays an important role in solving the complex problem of information. The research of text classification began in the late 1950s. It has experienced many different stages, such as the research of text classification theory, text classification technology based on expert system and text classification technology based on machine learning[1]. Now it has been widely used in the fields of search engine, information filtering, network forum and so on. Therefore, automatic text classification technology has great practical value and has made a lot of research progress, but there are still some deficiencies and difficulties, which are worthy of further discussion. Based on the research on the technology and algorithm of each stage of text classification, this paper proposes a text classification algorithm based on the mixed characteristics of word level and character level, collects news texts on the Internet by using crawler technology, designs and implements a news text classification system that can automatically classify and label, and can permanently store text classification information[2]. Applying the improved text classification algorithm to news text classification system has important research value and significance.

2 Model method

Compared with voice and image, text is more complex and abstract. Human beings can have an overall understanding of the text content after reading the text according to their own understanding ability. However, the semantics in natural language is difficult to be directly understood by computers. Therefore, the text content must be expressed as forms that computers can understand and process, such as 0 and 1. Text representation model is to use numerical or symbolic vectors that can be expressed by computer to represent abstract and complex natural text. In order to better represent the text, it is necessary to extract the most representative features from the text data[3]. These features should have obvious statistical laws, which can reflect the text distribution in the feature space and minimize the computational complexity of text mapping to the feature space[4].

2.1 Convolutional neural network

Convolutional neural network is a kind of feedforward neural network with convolution calculation and depth structure. It is one of the classical representative algorithms of depth learning. It has the ability of representation learning. The most important characteristics are "local perception" and "parameter sharing". It can classify the input information according to its hierarchical structure[5]. It is also called "translation invariant artificial neural network". As early as the 1980s, scholars began to study convolutional neural networks. The earliest ones were time-delay networks and lenet-5[6]. With the development of science and technology, the computing power of hardware becomes stronger, and convolutional neural network has developed rapidly. It was first applied to the field of machine vision, and later used to solve the problems in the field of natural language processing, and achieved good results in the task of text classification[7]. The structure of convolutional neural network is shown in Figure 2.

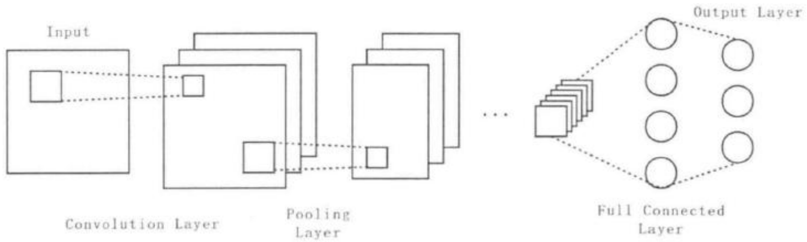

Fig. 1. The structure of convolutional neural network

Convolutional neural network is mainly composed of input layer, convolution layer, pooling layer, full connection layer and output layer. The input layer is used to receive input and process multi-dimensional data; The convolution layer is used for feature extraction, and there are multiple convolution cores in it; The pooling layer is used for feature selection and information filtering, which can reduce the output dimension of the convolution layer; The full connection layer is similar to the hidden layer of the traditional feedforward neural network. It combines the output of the pool layer nonlinearly; The output layer is the same as that of the traditional feedforward neural network. In the classification problem, the logic function or normalized exponential function is usually used to output the classification label.

2.2 Recurrent Neural Network

Recurrent neural network is a recurrent neural network that takes serialized data as input, recurses in the time direction of the sequence, and all computer nodes (cyclic units) are connected in a chain way. It has the characteristics of memory and parameter sharing. Therefore, it has certain advantages in learning the nonlinear features in serialized data, It is very suitable for dealing with video, voice, text and other timing related problems. RNN is called cyclic neural network because the output of each time in the sequence is related to the input of the current time and the output of the previous time. In order to transmit information, each node unit of RNN needs to be connected to form a recursive structure. The specific manifestation is: in addition to the input at the current time, RNN will also remember the previous information and apply it to the calculation of the current output, that is, the nodes between the hidden layers are no longer connected, but connected. The input of the hidden layer includes not only the output of the input layer, but also the output of the hidden layer at the previous time.

Although RNN can deal with timing related problems well, there are still problems such as gradient disappearance and gradient explosion. Sigmoid function is the most used activation function in neural networks. After the sigmoid function, the number from negative infinity to positive infinity is mapped between 0 and 1. When the neural network back propagates the error, it multiplies the partial derivative of the function layer by layer. Therefore, if the number of layers of the neural network is very deep, the deviation generated by the last layer will become smaller and smaller because it multiplies a lot of numbers less than 1, or even become 0, resulting in the weight of the shallow network layer not updated, which is the disappearance of the gradient. In addition, if the initial weight is too large, the weight of the hidden layer near the input layer changes faster than that near the output layer, which will cause the problem of gradient explosion.

2.3 LSTM

LSTM and Gru have two main methods to deal with gradient disappearance and gradient explosion: memorizing and forgetting information in a special way; Gradient clipping - when the gradient exceeds or is less than the threshold, the gradient at this time is set as the threshold. On the basis of RNN, LSTM adds a gate structure to delete or add information to the cell state. The structure of the gate is not complex, only a sigmoid layer and a point multiplication operation are combined. There are three gates in LSTM: input gate, forget gate and output gate. The input gate determines how much information enters the cell unit, the forget gate determines how much past cell information is discarded, and the output gate determines how much cell information is output. Figure 2 shows the network structure of LSTM.

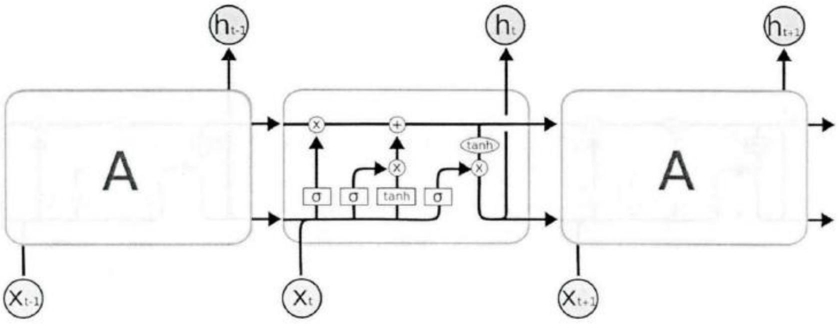

Fig. 2. The structure of LSTM

First, LSTM calculates the information to be discarded by forgetting the gate. The forgetting gate determines an output vector by observing the hidden state of the previous time and the input information X of the current time.Then, the input gate is used to calculate what new information needs to be added to the cell state. The input gate determines the information to be updated by observing the hidden state of the previous time and the input information of the current time. Similarly, tanh activation is performed on the two information to obtain the candidate cell information that may be updated to the current cell information.Then, by selectively forgetting the old cell information and inputting the candidate cell information obtained by selectively adding the gate, the old cell information is updated to obtain the cell information at the current time.Finally, the hidden state of the current time is determined by filtering the hidden state of the previous time and the input of the current time. Similar to the forget gate and the input gate, the output gate obtains a vector with all values between 0 and 1 for filtering information. The tanh layer obtains a vector with a value between - 1 and 1, multiplies this vector with the vector of screening information obtained by the output gate, and calculates the output of the final LSTM unit.

3 Experimental results and analysis

3.1 Experimental environment

The experimental environment of this paper is that the deep learning framework is pytoch, version 1.6, and the programming language is Python 3.6. The graphics card used is geforce-rtx960.

3.2 Dataset

The experiment uses two network news text data sets: Agnews and chnews / where Agnews is a public data set, Chnews is a data set with higher timeliness constructed by collecting Chinese news texts on China News Network (www.chinanews. Com) from January 1 to February 15, 2019. The two data sets are described in detail below.

Agnews is a search engine cometo myhead. Since 2004, Agnews has collected more than one million news articles from more than 2000 different news sources in more than a year, and the language is English. This experiment uses the subject classification dataset extracted and constructed by Zhang et al. The data set includes four categories: world, sport, business and SCI / TEC. Each category includes 30000 training samples and 1900 test samples.

News text dataset chnews, in Chinese. There are 21250 texts in the data set, including six categories: culture, international, social, financial, domestic and sports. In this experiment, 15000 pieces of text data were randomly selected as training data and 6250 pieces as test data.

3.3 Model training and results

In the part of convolution neural network, character based CNN is better than CNN using words as the basic unit. The reason is that the text length in the two data sets used in this paper is short. If words are convoluted directly, the information between words and adjacent characters will be omitted. In addition, because single-layer CNN can only obtain context information with fixed window size, the text information can be transmitted further by deepening the number of layers of CNN. Therefore, the accuracy of bilstm is higher than that of simple CNN model. However, deepening the layers of CNN will require higher computing resources. In order to provide users with more timely feedback, it is necessary to choose between the classification effect and the calculation time of the model. By adding attention mechanism before convolution layer, the model can make better use of the key information in the text and improve the accuracy of classification.

Tab. 1. Classification results

	Agnews	Chnews
CNN	86.16%	60.21%
LSTM	87.32%	62.31%
BiLSTM	89.23%	62.96%

4 Conclusion

This paper mainly focuses on the research of classification model based on deep learning and the implementation of classification system. By investigating the research status of text classification at home and abroad, and combined with the current cutting-edge research in the field of deep learning, this paper explores the classification model. The goal of this paper is to build a more accurate text classification model for network news, and judge which category label it belongs to for a given network news text. Based on this research goal, this paper uses bilstm and combines it with deep learning text classification method to construct two models, facn and FabG. Experiments are carried out on English data sets (Agnews) and Chinese data sets (chnews), which verify the effectiveness of the two models.

References

1. Mccallum A , Nigam K . A comparison of event models for Naive Bayes text classification. IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION, 1998:41--48.
2. Forman G . An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 2003, 3(2):1289-1305.
3. Baker L D , Mccallum A K . Distributional clustering of words for text classification. 1998:96-103.
4. Aggarwal C C , Zhai C X . A Survey of Text Classification Algorithms. Springer, 2012.
5. Zhao X , Kai Y , Tresp V , et al. Representative Sampling for Text Classification Using Support Vector Machines. Proceedings of the 25th European conference on IR research. Springer-Verlag, 2003.
6. F Peng, D Schuurmans. Combining Naive Bayes and n-Gram Language Models for Text Classification. Springer Berlin Heidelberg, 2003.
7. Chai K , Chieu H , Ng H T . Bayesian online classifiers for text classification and filtering. The Twenty-Fifth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. DSO National Laboratories 20 Science Park Drive Singapore 118230, 2002.