Open Access Library Journal
2021, Volume 8, e7175
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Review of Text Classification in Deep Learning

Qi Wang[1], Wenling Li[1], Zhezhi Jin[2]*

[1]College of Science, Yanbian University, Yanji, China
[2]Department of Economics and Management of Yanbian University, Yanji, China
Email: wangqi675@sina.com, *jinzhezhi@sina.com

How to cite this paper: Wang, Q., Li, W.L. and Jin, Z.Z. (2021) Review of Text Classification in Deep Learning. *Open Access Library Journal*, **8**: e7175.
https://doi.org/10.4236/oalib.1107175

Received: January 22, 2021
Accepted: March 22, 2021
Published: March 25, 2021

Abstract

Text classification is an important research content in natural language processing. Compared with traditional manual processing, text classification based on deep learning improves both efficiency and accuracy. However, in the learning process, the content involved is very large and complex. In order to facilitate the research of more scholars, this paper summarizes the text classification of deep learning. The first part of this paper introduces the pre-processing of text classification. The second part introduces several feasible methods for deep learning text classification in detail. The third part introduces the test method of the model. The fourth part summarizes and analyzes the advantages and disadvantages of several methods to lay a foundation for further research.

Subject Areas

Information Science

Keywords

Text Classification, Word-Embedding, Convolution Neural Network, Long and Short Term Memory Network

1. Introduction

Natural language processing is to use computers to process and use human language, it is a hot topic in the field of artificial intelligence, and has brought great convenience to our life and study, including common machine translation, speech recognition, public opinion analysis, text classification and so on. These involve data mining, machine learning, artificial intelligence and many other researches. In these researches, text classification is the basis of other researches and plays a very important role.

*Corresponding author.

Previously, text classification can be divided into three categories, the first is the traditional manual text classification. The main problems of this research method are as follows: 1) the traditional text classification will waste a lot of manpower and material resources; 2) artificial text classification will produce great errors. The second type is based on traditional shallow machine learning methods, such as support vector machine (SVM) [1], Naive Bayes [2], decision tree [3], K Nearest Neighbor [4], maximum entropy [5] and so on. The main problems of this research method are as follows: 1) the processed text generally has the disadvantage of high dimension and sparse vector, which is easy to produce gradient explosion or gradient disappearance in the follow-up research; 2) shallow machine learning methods are time-consuming and difficult to operate. The third category is text classification methods based on deep learning. Generally, convolutional neural network and recurrent neural network are generally used. This method was initially applied in the field of deep learning image processing, and gradually applied to the field of text with the further development of research. At present, this kind of method is the most widely used, but the problems of this kind of research method are as follows: 1) there are many kinds of methods and each method has its own advantages and disadvantages, so it is not suitable to choose in practical problems; 2) if handled improperly, problems such as gradient disappearance or gradient explosion will occur.

In the practical problems of text classification, the third kind of research method can not only save time and effort but also improve accuracy. Therefore, this paper will solve the problem of deep learning text classification and divide it into the following research tasks: 1) pre-processing of text classification; 2) summarize the characteristics of existing neural network models, and compare the advantages and disadvantages of each model; 3) the method of testing the model is introduced; 4) summarize the problems existing in the current deep learning text classification research, and put forward the prospect of the follow-up research.

2. Preprocessing of Text Classification

Text classification should be done by transforming human language into computer language, that is, converting text into vector form that computer can recognize. There are some problems in text pre-processing at present: 1) Text size is different, which will lead to different vector dimension; 2) chinese text is different from English text, and there is no space distinction between word phrases; 3) there are a lot of meaningless modal words in the text; 4) the traditional one-hot coding has sparse vector and high dimension, which is not easy to follow up processing. Based on the above problems, this paper will introduce a better conversion mode.

2.1. Participle Treatment

Large text cannot be directly converted into vectors, so the first step in text preprocessing is to segment the text or sentence into phrases. “Jieba” segmentation is a common word segmentation tool.

There are 3 modes of "Jieba" segmentation, namely precise mode, full mode and search engine mode. Precise mode is the most accurate one, which is the most suitable one for text classification. Stuttering segmentation provides keyword extraction and custom dictionary functions, which can extract some special unrecognized proper nouns and add them by themselves. Stuttering segmentation can be a good way to cut out custom words. In addition, there are many modal words in the text, they have no practical meaning, so the stuttering segmentation also provides this function.

2.2. Word-Embedding

Once text is processed into phrase form, it needs to be converted to vector form, a mapping process also known as "word-embedding." The traditional method of word embedding is usually one-hot coding, but this method takes character as the unit, and the vector is sparse, which can cause gradient explosion or gradient disappearance. To solve this problem, word2vec is proposed. Word2vec consists of two models, namely, continuous bag of words (CBOW) and skip-gram, which predicts the current word in context and skip-gram, which predicts the context with the current word. CBOW is suitable for dealing with small text data, and text classification problems are mostly long text, so skip-gram is more commonly used, which performs better in large corpora.

3. Introduction of Text Classification Problem Based on Deep Learning

3.1. Convolution Neural Network (CNN)

Convolution Neural Network is initially applied to image processing in deep learning. CNN can capture important parts of images, and has strong ability to extract and share weights. In 2015, Kim [6] proposed applying CNN to text processing and segmenting sentences. Convolution Neural Network is a kind of common deep learning network, which is composed of input layer, convolution layer, pooling layer, full connection layer and output layer. Structure diagram like Figure 1.

Figure 1 shows a general convolution neural network structure. The process of feature extraction is carried out in convolution layer and pooling layer. In practice, the parameters of convolution layer and pooling layer can be changed according to specific requirements.

The convolution layer is composed of many convolution kernels, in which the convolution kernels are a kind of filter and matrix with weights. Then, the convolution operation of the i window can be expressed as:

$$Y_i = g\left(x_i \cdot W + b\right) \tag{1}$$

where, Y_i represents the i-th vector of the convolution output; x_i represents the input vector in the i-th window; W represents the weight of each vector; b is the offset term.

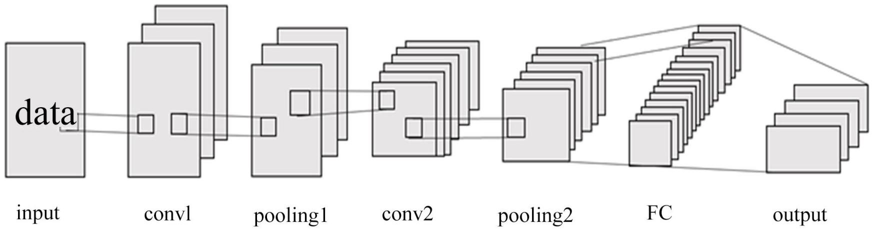

Figure 1. General convolution neural network diagram.

Pooling layer is also composed of filters, pooling layer can be divided into maximum pooling, average pooling, minimum pooling, which means that the vectors of the previous convolution layer are maximized, averaged, and minimized. In word processing, the most common method is maximum pooling, which is to extract the most obvious features of vectors. Maximum pooling operations can be expressed as:

$$v[i] = \max Y[i] \tag{2}$$

where, $v[i]$ represents the vector after the pooling operation at layer i; $Y[i]$ represents the vector after the i-th convolution operation.

Maximum pooling can select the maximum value of each layer, but in practical application, if only one maximum value is selected, some other important information will be lost. Thus, K-max pooling is extended. K-max pooling means to select K maximum values in each layer, which is commonly used in text classification processing.

The full connection layer plays the role of "classifier" in the whole network and splice all the pooled vectors together. In the application of text classification, "softmax" function is usually added after the full connection layer to carry out predictive classification.

The operation of convolution neural network can extract the important content of the article and share the weights, which is also the advantage of convolution neural network in image processing. But the meaning of some words in the text is different from the picture, and the meaning of some words in the text needs to be determined by context, while the convolution neural network can not convey much information and is not enough to contact the context. Therefore, we need to understand the recurrent neural network.

3.2. Recurrent Neural Network (RNN)

Traditional neural networks can't predict the information of the next node based on the information of the previous node, and the words in each sentence are not independent in the text. The greatest advantage of recurrent neural network is that they can transmit the previous information to the later, that is, the current output of a sequence is related to the previous output. A simple diagram of the structure of the RNN is as **Figure 2**.

The right side of the figure is an expansion of the left side, with information from the previous for each time t, controlled by a weight of W. However, it can be seen from the picture that each front node has nothing to do with the following one. Therefore, M. Schuster and K. Paliwal proposed a bidirectional Recurrent neural network (BRNN) [8], and the simple structure is shown as **Figure 3**.

The bidirectional Recurrent neural network can connect context semantics well in text processing, which makes text analysis more accurate. But BRNN also has the problem of long-term dependence on RNN. Most of the text in the text categorization application are long articles. Every node in BRNN network model is related to each other, so the data volume is too large and time consuming may be long. To solve this problem, another branch of RNN, the Long and short term memory networks (LSTM), is introduced.

The long and short term memory network model consists of input gate, forgetting gate and output gate. The basic structure diagram is like **Figure 4**.

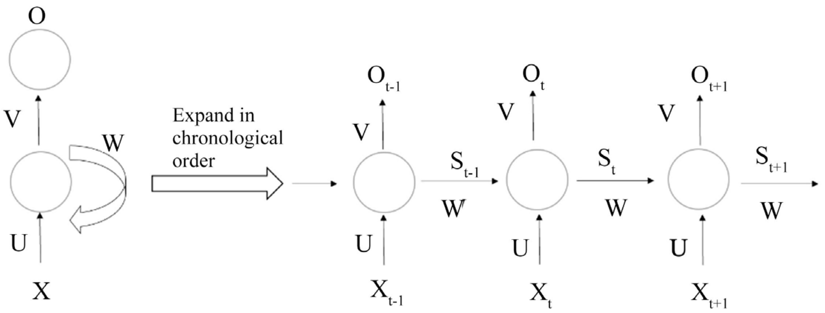

Figure 2. RNN structure based on time. (Note: picture reference [7]).

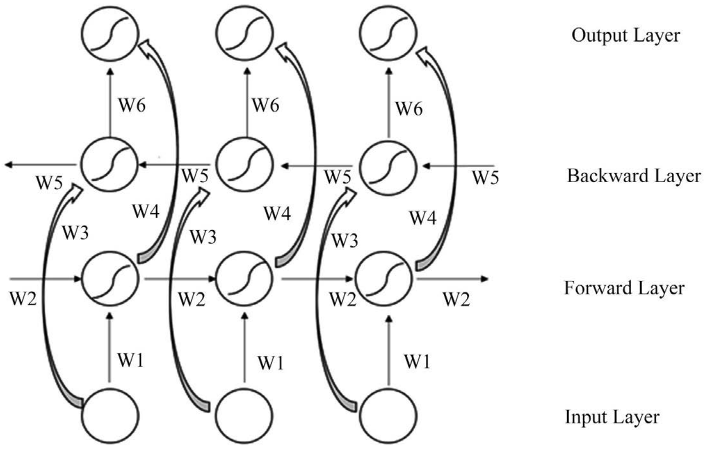

Figure 3. BRNN simple structure chart. (Note: picture reference [8]).

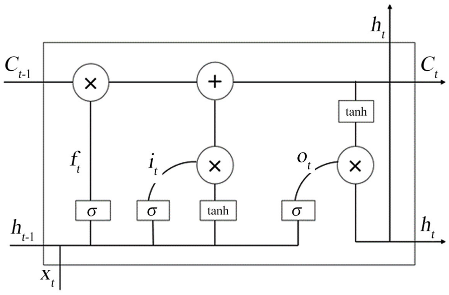

Figure 4. LSTM simple structure diagram. (Note: picture reference [9]).

In the figure, storage unit C_t is the historical information of the current moment; C_t is the input gate, which determines the change amount of the input vector to the stored information at the current moment. f_t is the forgetting gate, which determines the influence degree of the previous moment on the current moment; O_t is the output gate, which controls the input vector at the current moment. Assuming that the input vector is $X = [x_1, x_2, x_3, \cdots, x_n]$, the calculation formula is as follows:

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{3}$$

$$O_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \tag{4}$$

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{5}$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{6}$$

$$h_t = O_t \otimes \tanh(C_t) \tag{7}$$

where h_t is the final output of LSTM; $\sigma(\bullet)$ is the sigmoid activation function; $\tanh(\bullet)$ is the tangent function; $\otimes$ A new vector is formed by the product of the corresponding elements of two vectors of the same dimension; $\otimes$ Add the corresponding elements of two vectors of the same dimension to form a new vector; W_i, W_o and W_f are the weights of input gate, output gate and forgetting gate respectively; b_i, b_o and b_f are the offset terms. Long and short term memory network model is a major breakthrough in the field of deep learning text. Through the control of the special structure "gate" in LSTM, important contents in the sequence can be retained, the redundant parts can be deleted, and then predictions can be made along the relevant information transmitted. Long and short term memory network solves the problem of long term dependence of RNN and is a popular network for text classification.

Along with the wide use of long and short term memory networks, many problems have been exposed, such as long training time, many parameters, complex internal calculation and so on. So some scholars put forward the GRU model, GRU model is a simplified LSTM model, GRU only has two "gates", namely "reset gate" and "update gate". GRU model has fewer parameters, so the training speed is improved than LSTM, and less data are needed in training. But if there are enough data, LSTM's powerful expression ability will produce better results. Which kind of network model should be chosen from practical problems.

4. Verification of Models

For trained classifier, it is necessary to test its generalization, inspection is generally will not processed the text into the trained classifier, will not processed text categorization, classification after compared with the category of the real, is obtained by the experimental results can be used the following criteria for evaluation, usually adopt evaluation index such as accuracy and recall rate, F_1 value, first set up the confusion matrix.

Table 1. Confusion matrix.

Predictive value Real value	Positive case	Negative cases
Positive case	TP (True Positive)	FN (False Negative)
Negative cases	FP (False Positive)	TN (True Negative)

The TP in Table 1 indicates that X samples are correctly assigned to Class X, FP indicates that other samples are wrongly assigned to Class X, FN indicates that X samples are wrongly assigned to Class X, and TN indicates that other samples are correctly assigned to Class X. From the above table you can calculate:

$$\text{Accuracy}(\text{ACC}) = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

$$\text{Error rate} = \frac{\text{FP}+\text{FN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{F}_1\text{-Measure}(\text{F}_1) = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}$$

Accuracy is the most common index, generally speaking, the higher the accuracy, the better the classifier. But in the special case, it is not enough to calculate the accuracy, recall and F_1 value, in which F_1 is a synthetical evaluation index, combined with the accuracy can better test whether the classifier is accurate or not. The larger the F_1 value, the better the fitting degree is.

5. Summary and Analysis

This paper introduces and summarizes the preprocessing of text classification, the related calculation methods and the test methods. The advantages and disadvantages of the related models are sorted out, in which the CNN model can capture the important content of the text, while the RNN model can analyze the context. The deep learning method is applied to text classification, which saves a lot of manpower and material resources and improves the accuracy of text classification. But the deep learning network models are various and each has its own advantages and disadvantages, so it is necessary to choose and use the model. The author will continue to study it in this direction.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Che, J.H., Feng, Y.X., Tan, J.R., *et al.* (2007) Classification of Product Design Knowledge Documents Based on Decision Support Vector Machine. *Computer In-*

tegrated Manufacturing Systems, No. 5, 891-897.

[2] Pang, X.L., Feng, Y.Q. and Jiang, W. (2008) Compensation Strategy for Missing Feature Words in Bayesian Text Classification. *Journal of Harbin Institute of Technology*, **40**, 956-960.

[3] Zhu, Y.P. and Dai, R.W. (2005) Text Classifier Based on SVM Decision Tree. *Pattern Recognition and Artificial Intelligence*, **18**, 412-416.

[4] Yu, Y., Miao, D.Q., Liu, C.H., *et al.* (2012) An Improved KNN Classification Algorithm Based on Variable Precision Rough Set. *Pattern Recognition and Artificial Intelligence*, **25**, 617-623.

[5] Huang, W.M. and Sun, Y.Q. (2017) Chinese Short Text Sentiment Analysis Based on Maximum Entropy. *Computer Engineering and Design*, **38**, 138-143.

[6] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, 1746-1751. arXiv: 1408.5882
https://doi.org/10.3115/v1/D14-1181

[7] Ji, L.K. (2019) Research on Text Sentiment Analysis Technology Based on Deep Learning. Beijing University of Posts and Telecommunications, Beijing.

[8] Schuster, M. and Paliwal, K. (1997) Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, **45**, 2673-2681.
https://doi.org/10.1109/78.650093

[9] Wang, H.T., Song, W. and Wang, H. (2020) A Text Classification Method Based on LSTM and CNN Hybrid Model. *Journal of Small and Microcomputer Systems*, **41**, 1163-1168.