Research on Chinese Text Classification Algorithm based on Convolutional Neural Network

Junchao Wei

Xi'an Traffic Engineering Institute, Xi'an 710300 China

Abstract. Various information in the era of Internet big data has shown an "explosive" growth, and mining useful information from text data information is one of natural language processing content. In addition to major breakthroughs in image recognition, deep learning convolutional neural networks can also be applied to text classification. Taking Chinese data as the research object, a new text classification model is constructed by using the CNN algorithm and the jump-gram combination of convolutional neural networks. At the same time, the traditional Pinyin classification methods are compared. Through simulation experiments, it is proved that the CNN algorithm has a good effect on text classification, and its classification accuracy is as high as 88%.

Keywords: Classification Algorithm; Chinese Text; Convolutional Neural Network.

1. Introduction

With the rapid development of Internet technology, we are in the era of information explosion, while enjoying the convenience brought by rich online information, we also face the severe challenge of how to extract data information quickly and efficiently from massive information. Among them, the text categorization (TC) technology [1-3], which plays a crucial role in the processing of massive text data, is mainly to solve the problem of disorganized text information to a certain extent. The basic principle is to prescribe the set of label labels. Extracting the relevant text features from the original text content to finally determine the category label to which it belongs, and with it a lot of applications and research.

As the key technology of information processing, text classification technology is an important foundation of information organization and text mining. It has been widely used in the fields of knowledge mining, information retrieval and information supervision. There are many methods for text categorization. The more classic ones are: Naive Bayesian classifier [4], K nearest neighbor (KNN, K-Nearest Neighbor) algorithm [5], support vector machine (SVM) [6] and Back. Propagation neural network [7] and other classification models, and have achieved good results. But in general, these shallow neural network algorithms face common limitations: local optimization, dimensionality disasters, over-fitting, etc. [8], and shallow networks in limited samples and computational units. The limited expression of complex functions in the case leads to the generalization of its ability to deal with complex classification problems.

Domestic research on text classification began in 1980. The main development process also includes the following three steps: classification theory knowledge research, knowledge engineering establishment expert system and automatic classification system based on machine learning. Taking the news text as an example, this paper studies the text classification problem, and uses the model algorithm based on convolutional neural network to realize the re-extraction of text features, including Chinese text and word features. Then design the convolutional neural network model and conduct experiments to find the optimal values of the relevant parameters.

2. Convolutional Neural Network

2.1 The Definition of Convolutional Neutral Network

The definition of a convolutional neural network is as follows [9]: The two sets of V_e and E_d are defined to form a convolutional neural network, which is recorded as:

$$G = (V_e,\ E_d) \tag{1}$$

In this type, E_d represents a finite set of edges, and V_e represents a non-empty finite set of nodes. When these edges have directions, they are called directional convolutional neural networks, and vice versa, called undirected convolutional neural networks.

By means of the similarity between the convolutional neural network structure and the group structure, and the adjacency matrix used does not need to know the distance between the texts, the difficulty can be greatly reduced, and a matrix of text adjacency can be obtained, as follows:

$$T_d = \begin{bmatrix} 0 & t(1,2) & t(1,n) \\ t(2,1) & 0 & t(2,n) \\ t(n,1) & t(n,2) & 0 \end{bmatrix} \tag{2}$$

In the type, Where t(i,j) is equal to 1 when the i-th text is the jth text parent node; in other cases, t(i,j) is equal to 0.

If the text has a single parent node, the text classification model is as follows:

$$x_{k+1,i} = F_{k,l}x_{k,i} + b_k(\mathrm{l}, \mathrm{i}) + B_{k,i}w_{k,i} \tag{3}$$

$$z_{k+1,i} = C_{k+1}x_{k+1} + v_{k+1,i} \tag{4}$$

Where $x_{k,i} = \left[P_{k,x}x_{k,i},\ v_{k,x}x_{k,i},\ P_{k,y}y_{k,i}, v_{k,y}y_{k,i}\right]T$, $P_{k,x}y_{k,i}$, $P_{k,y}y_{k,i}$, $v_{k,x}x_{k,i}$ and $v_{k,y}y_{k,i}$ represents the position and velocity of text i on the x and y axes, respectively; $x_{k,i} \in x_k$. l represents the parent of the i text. b_k(l,i) is a compensation vector representing the positional relationship between text i and its parent node; F and C respectively represent the word vector transfer matrix and observation matrix; B is the word vector noise figure matrix; w and v are volumes the neural network noise respectivelyand observed noise are subject to a normal distribution.

2.2 Convolutional Neural Network Structure

The convolutional neural network mainly consists of three layers: convolutional layer, pooled layer, fully connected + Softmax layer.

2.2.1 Convolutional Layer

Each layer of convolutional layer is composed of several convolutional units, and the parameters of each convolution unit are optimized by a back propagation algorithm. The purpose of the convolution operation is to extract different features of the input. The first layer of convolutional layer may only extract some low-level features such as edges, lines and corners. More layers of networks can iteratively extract more complex from low-level features. feature.

2.2.2 Pooling Layer

The pooling layer is also called Down sampling, and the opposite is Up sampling. It is mainly used for feature dimension reduction, compressing the number of data and parameters, reducing over-fitting, and improving the fault tolerance and training speed of the model. There are two ways to sample: Max Pooling and Mean Pooling.

2.2.3 Full Connected + Softmax Layer

After multi-layer convolution and pooling operations, the obtained feature maps are sequentially expanded in rows, connected into vectors, and input into a fully connected network. Softmax logistic regression is usually used as the feature classifier.

2.3 Characteristics of Convolutional Neural Networks

Convolutional neural networks include three categories of features, first of all local perception, which means that the nodes of the convolutional layer are only connected to some of the nodes of the

previous layer, used to learn local features, and the way of this connection greatly reduces the number of parameters. To speed up the learning efficiency and reduce the possibility of over-fitting to a certain extent. The second is the spatial arrangement, that is, the size of the output unit is controlled by the following three quantities, respectively depth, which controls the output unit depth (the number of filters) and the neurons connecting the same block area.

Stride controls the distance between two adjacent hidden units at the same depth and the input area connected to them; zero-padding controls the overall size of the input unit by zero-padding around the input unit, thereby controlling the space of the output unit. size. The third type of feature is weight sharing, that is, the plane of the same depth is called the depth slice, the same slice shares the same set of weights and offsets, and the repeating unit can identify the feature without considering its position in the visible domain. This helps the neural network maintain spatial invariance on the input.

3. Text Classification Algorithm

The core formula of Naive Bayes is as follows:

$$P(B|A) = \frac{P(B|A)}{P(A)} \tag{5}$$

The main idea is as follows: It is assumed that the item x to be classified is composed of the feature items $\{t_1, t_2, ..., t_m\}$, and the probability of each category y_i appearing by calculating the occurrence of each feature item, t_i, it is taken, and the category of the maximum probability is taken. As the category of item x to be classified. That is, after calculating $P(y1|x), P(y2|x), \ldots P(yn|x)$ in turn, do the following values:

$$P(yk|x) = \max\{P(y1|x), P(y2|x), \ldots, P(yk|x)\} \tag{6}$$

The category of the classified item is finally determined. Under the assumption that the characteristics of each feature are independent, the formula of $P(y_i|x)$ is as follows:

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)} = \frac{P(t_1|y_i)P(t_2|y_i)\ldots P(t_j|y_i)}{P(x)} \tag{7}$$

4. Chinese Representation Method and Character Level CNN

The representation of Chinese text is usually divided into two types, namely word level and character level representation. The word level has only one representation because of its unique reason, and the character level has different coding representation methods, including pinyin coding, UTF-8 coding, picture coding, and random embedding vector methods. Based on the random embedded vector, a method of pre-training character embedding vectors is proposed. The idea of this method is consistent with the idea of distributed expression, that is, Chinese characters with the same context should have similar vector representations.

4.1 Phase Level Representation

When the Chinese text adopts the phase-level text representation method, the text needs to be processed by word segmentation, and then denoised to obtain all the feature phases of the style, and the feature words are stored in a dictionary. In general, there are three steps to denoise: the first step is to find the frequency of occurrence is less than the specified value, and the second step is to create a stop word library. The third step is to replace the representation of the text data using the index of the feature vocabulary after establishing the feature dictionary of the data set. After learning the embedding vector of all feature words, a piece of text can be represented as a matrix.

4.2 Word Level Representation

There are usually many ways to express text in Chinese character level. The commonly used pinyin coding method requires word segmentation and then conversion to pinyin format. Another way is UTF-8 encoding. Because all Chinese characters can be expressed in UTF-8 encoding, as summarized in Table 1. u indicates the start bit of the character encoding, and the letter after the u and the number indicate the hexadecimal number. It can be seen from the table that a Chinese character usually requires a 5-bit UTF-8 encoding representation. Some special characters require 6 digits, so when the Chinese text is expressed in UTF-8 encoding, the length of the text will be more than 5 times the original length. As with the Pinyin format, the data length of the original text is expanded. A larger convolution kernel width or a deeper convolutional network is required for convolution to extract more abstract conceptual features.

Table 1. Pinyin, UTF-8 and word vector representation of Chinese characters

Character	Pinyin	UTF-8
公	gong1	u516c
小鸟	xiao3niao3	u5c0fu9e1f
密码	mi4ma3	u5bc6u7801

5. Character Level CNN Model Design and Test Results

In the feature extraction problem of Chinese text, the concept of word vector is introduced. Based on the Skip-gram language model model provided by Word2vec, the relationship between the features of the text is presented in the form of a word vector. To put it another way, words with very similar semantic understandings have a particularly close distance in the embedded space. Thereby achieving a seamless conversion of word granularity and sentence granularity in the text, and finally obtaining a word vector representation of the text. The Skip-Gram mode is shown in Figure 1. Directly from the diagram, the model is known to be the current word, and then the probability of occurrence of the word before and after the current prediction is performed.

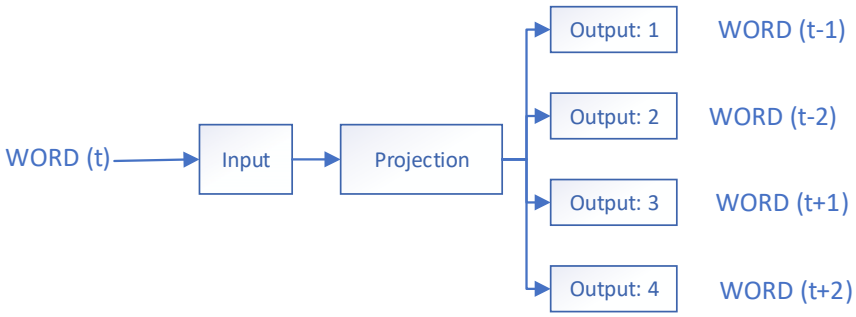

Fig.1. The Skip-gram model structure

According to the skip-gram model design principle, according to the experimental environment and configuration of a certain Chinese text, the results obtained after comparison with the Pinyin representative method are shown in Table 2:

Table 2. Comparison of experimental results

Model	Accuracy %
CNN+skip-gram	88
Pinyin	69

It is not difficult to find from the above table that the CNN + Skip-gram model has the highest classification accuracy in the text classification problem, and its accuracy rate reaches 88%. This proves that the classification of news texts based on CNN + Skip-gram model is more significant.

6. Conclusion

Aiming at the shortcomings of neural network classification algorithm in text classification, a text classification algorithm based on convolutional neural network is proposed. By using the CNN + skip-gram model combination, and then using the classification word vector of each text at each moment, the adjacency matrix classification can be obtained every time, and the adjacency matrix classification can be used to obtain the subgroup number classification, and the classification is accurate. The rate is as high as 88%. Simulation experiments show that the classification effect of CNN algorithm in text is more significant.

Acknowledgements

2017 Annual Scientific Research Projects of Shaanxi Association of Higher Education: "Research on Mathematical Knowledge and Computer Skills for Students' Career Development in Application-Oriented Institutes"(No.XGH17246); 2018 Funded Projects for Middle-Aged and Young Teachers of Xi'an Traffic Engineering Institute: "Design and Implementation of Experiment Platform of Higher Mathematics Based on MATLAB" (No.KY18-40).

References

[1]. Zhou Y, Hong C, Zhu Q. The research of classification algorithm based on fuzzy clustering and neural network[C]. IEEE International Geoscience & Remote Sensing Symposium. 2002.

[2]. Ren X, Yi Z, He J, et al. A Convolutional Neural Network-Based Chinese Text Detection Algorithm via Text Structure Modeling[J]. IEEE Transactions on Multimedia, 2017, 19(3):506-518.

[3]. Du J H. Automatic text classification algorithm based on Gauss improved convolutional neural network[J]. Journal of Computational Science, 2017, 21: S1877750317307238.

[4]. Manupati V K, Akhtar M D, Varela M L R, et al. A Text Mining Based Supervised Learning Algorithm for Classification of Manufacturing Suppliers[J]. 2018.

[5]. Khan S, Baig A R. Ant colony optimization based hierarchical multi-label classification algorithm[J]. Applied Soft Computing, 2017, 55:462-479.

[6]. Zhang X, Zhou X, Lin M, et al. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices[J]. 2017.

[7]. Zhang H, Ying L, Zhang Y, et al. Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network[J]. Remote Sensing Letters, 2017, 8(5):438-447.

[8]. Sudholt S, Fink G A. PHOCNet: A Deep Convolutional Neural Network for Word Spotting in Handwritten Documents[C]. International Conference on Frontiers in Handwriting Recognition. 2017.

[9]. Sahiner B, Chan H P, Petrick N, et al. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images[J]. IEEE Trans Med Imaging, 1996, 15(5):598-610.