Commented Content Classification with Deep Neural Network Based on Attention Mechanism

Qinlu Zhao[1] ,Xiaodong Cai[1]* ,Chaocun Chen[1] ,Lu Lv[1] ,Mingyao Chen[2]
1.School of Information and Communication,Guilin University of Electronic Technology,China
2.Guilin Topintelligent Communication Technology Co.,Ltd,China
*Corresponding author: caixiaodong@guet.edu.cn

Abstract— **It is difficult to fully represent text information with shallow network, and it is time-consuming for using deep neural network. This paper proposes a CNN-Attention network based on Convolutional Neural Network with Attention (CNNA) mechanism. First of all, information between words for context can be expressed by using different sizes of convolution kernels. Secondly, an attention layer is added to convolution network to obtain semantic codes which include the attention probability distribution of input text sequences. Furthermore, weights of text representing information are calculated. Finally, the softmax is used to classify emotional sentences. Experimental results show that features of different context information can be extracted by the method proposed, the depth of the network is reduced and the accuracy effectively is improved at the same time. It also shows improved accuracy in COAE2014 task 4 micro-blog data set for emotional classification up to 95.15%.**

Keywords—deep neural network; CNN; attention mechanism; emotional sentences

I. Introduction

In recent years, deep learning is widely used in natural language processing. Research on emotion analysis has attracted large number of researchers. Highly remarkable results in sentence classification task are achieved by CNN[1~3]. In [2], input data is one-hot high-dimensional vector used to represent word vector as input in CNN. Some recent works shown that word vector obtained by using unsupervised learning can greatly improve the accuracy of the model in [4~6]. In[7], word vectors are learned by predicting the context information of words using the Skip-Gram model. The words with similar semantics can be clustered by the model in the word vector space.

CNN is an alternative network with the characteristics of weight sharing and local sensing. It consists of convolutional layer and sampling layer. In [8], local information is processed by employing convolution kernel filter of each layer with CNN. In [1], the feature vectors in sentence level are learned by different convolution kernels. Then the deep information of the text is acquired by connecting these feature vectors. Curse of dimensionality can be avoided by using the method of CNN. Attention mechanism which focuses on a particular area and ignores the other parts at some time is a model of brain resource allocation . It is applied for image processing and natural language processing. In [9],Global and Local Attention models are used in machine translation. In [10], Attention Model is utilized for text auto-summarization. In [11], the pooling network based on the attention mechanism is employed in a question answering system. The semantic codes including the attention probability distribution of the input text sequences are obtained by the method of attention mechanism. Furthermore, weights of text representing information are calculated.

II. Classification Based on CNN Attention for emotional polarity

There are four parts of the proposed network model for emotional polarity classification: the vector representation of the text, the sentence presentation layer, the text feature extraction and the text categorizer. The framework is shown in Fig.1.

A. Word vector initialization

In this paper, the Skip-Gram unsupervised model in word2vec [12]is adopt for vectors training. The word vectors contain words appearing more than three times in corpus. Otherwise, the words are vectorized with the mean vector generated by word2vec. It is presented as $M \in R^{dx|V|}$, where d is the latitude of the word vector and V is total number of the words in the corpus. The i-th word in the sentence is expressed by $X_i \in R^{dx1}$. In this work, each sentence is unified into n words and it can be written as $S_{1:n}=\{s_1,s_2,s_3,......,s_n\}$, and the sentence sequence S is denoted as $S \in R^{dxn}$ of 2D data.

B. Text feature extraction

1) CNN context information extraction layer

Different context information of sentence is extracted by convolving with different size of kernels. In this work, the text sequence $S_{1:n} \in R^{dxn}$ is convolved with different sizes of convolution kernels which are $W_1 \in R^{1x1}$ 、 $W_2 \in R^{1x2}$ 、 $W_3 \in R^{1x3}$ 、 $W_4 \in R^{1x4}$ 、 $W_5 \in R^{1x5}$. The stride of convolution is 1.Context information of words is obtained and multilevel representation of text information is extracted. The convolutional method can be described by Eq. (1):

$$c_i = sigmoid(w_i \otimes x + b) \qquad (1)$$

Where W_i is a convolution kernel which contains i words. X is the input sequence information following vectorization. Sigmiod is non-linear activation function.

In order to enhance the expression ability of network model in text input sequences. The nonlinear activation function sigmoid is used to normalize the output of the convolution layer. Different feature maps of context, such as $C1 \in R^{dxn}$, $C2 \in R^{dx(n-1)}$, $C3 \in R^{dx(n-2)}$, $C4 \in R^{dx(n-3)}$, $C5 \in R^{dx(n-4)}$, are obtained by this method. Then feature maps are operated with max pooling. The pooling kernel is set to

978-1-4673-8979-2/17/$31.00 ©2017 IEEE

2X2 and stride is *2* in the network. The context information is further extracted and the size of feature map is reduced.

2) Extraction Model of Attention Mechanism Information

Attention probability distribution matrix of the input text sequence is obtained by attention mechanism. The weight of text feature information is calculated by the matrix, which reduces the missing and redundant information during feature extracting.

The feature maps of different context words information are obtained by output of pooling layer. Attention mechanism is operated using dot product of attention matrix and different maps. Attention matrix is generated randomly. Then the matrix parameters are optimized by feedback training with a large number of iterations. Finally, the attention matrix is optimized. The importancy of region information is determined by different values of matrix which represents the weight of word information. Furthermore, more complex features are extracted with the matrix. Pseudo code is shown in Table I.

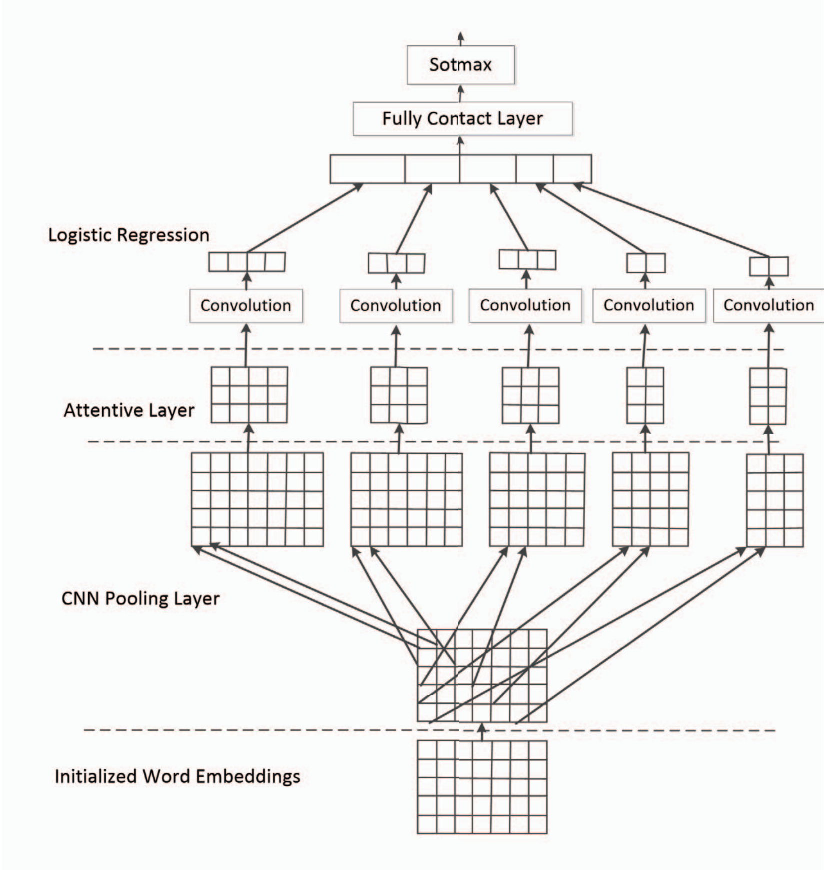

Fig. 1. The framework of the Cnn-Attention network

TABLE I. PSEUDO-CODE OF ATTENTION MECHANISM

```
input：$T=T_{im}$，eg.i=0，1，2.，3，4 ; m = 0,1,…,15    Attention={attention|i=1,2,3,4,5}
output：Att_splitMap
Begin
Initialize $mat_i$=random()，
            Where i = 0，1，2，3，4，5 ; /* All attention matrixes are randomly initialized */
 Attent_$mat_{i=}$softmax($mat_i$*tanh($W_w$*$mat_{i+}h_i$))
For m=1 to m
   For i =1 to i
      Split_map=split_Feature_Map(T) ; /* Split the feature map matrix */
      Att_Split_Map=Split_Map; /* dot multiplication Attention Matrixes */
      Att_Split_Map += Split_Map;/* Merge feature map matrix */
   Return Att_split_Map/* Returns the new feature map matrix */
End
```

C. Sentiment classification

Feature maps considering different words in context are extracted by neural networks and used as the feature of input sequences. The probability distribution of these classes is obtained by softmax classifier and is presented by Eq. (2):

$$y_i = \frac{\exp(x_i)}{\sum_{j=0}^{n-1} \exp(x_j)} \qquad (2)$$

Where, x_i is the i-th node value of softmax layer, y_i is the i-th output value, n is the node number of softmax layer.

III. THE EXPERIMENTAL RESULTS AND ANALYSES

A. Network training

COAE2014 task 4 weibo data set is used in the experiment. The total number of data is 40000. 5000 weibo records of emotional polarity are released in the data set. Each sentence is unified into 40 words and word2vec tools is used to implement word vector operations. Words not appearing in the word vector are replaced by average value of word vectors produced by word2vec.The parameters used for training are shown in Table II. In this paper, experimental platform equipped with Intel i5-4460 (4x3.4GHz) CPU, 8GB RAM, GTX750Ti graphics and Ubuntu14.04 operation system. Tensorflow toolkit based on C++ programming language is used to train network in this paper.

TABLE II. ADJUSTABLE PARAMETER SETTINGS OF WORD2VEC WORD VECTOR

Tunable parameters	parameter values
Word vector dimension	20、50、100、200、300
Word vector dimension	1
Algorithm selected	Skip-Gram
Context window	10
Sampled value	1e-3

Supervised training is adopt in this work. Each sample of input sentences for training is labeled into different emotional categories. A cross-entropy loss function is used to train the model. The dimensions of word vectors are 20, 50, 100, 200 and 300 in this experiment. The length of sentence is 40. The experimental model is utilized with the COAE2014 task 4 corpora. Then the best experimental result is selected after numbers of iterations. Different dimensions of word vectors are used in the network and the results are shown in Fig.2:

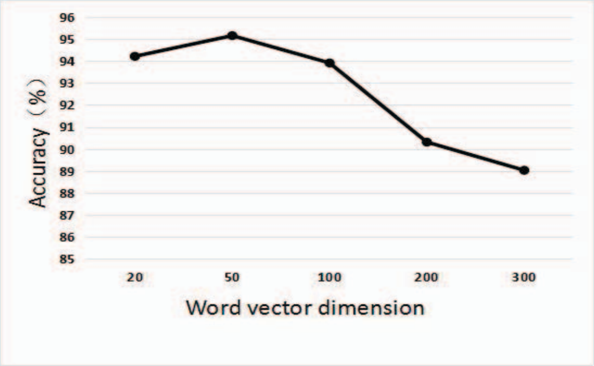

Fig. 2. Accuracy adapting to different dimensions

In order to verify the effectiveness of proposed method, SVM and CNN-based methods are compared. The experimental results as shown in Table III:

TABLE III. DESCRIPTION OF MODELS

SVM bow1	Vector features is unigram. They are classified by SVM.
SVM bow2	Vector features are unigram and bigram . They are classified by SVM .
SVM bow3	Vector features are unigram and bigram and trigram. They are classified by SVM .
CNN-word-static	A model with pre-trained vectors from word2vec. All words including the unknown ones that are randomly initialized are kept static and only the other parameters of the model are learned.
CNN-Attentive	Using word2vec to train the secondary level word vector experiments. The word vector is not fine-tuned during the experiment, but the CNN and the attention matrix parameters are learned.

The experimental results show that the accuracy achieves 95.15% when the word vector is 50-dimension. The comparison of our method and other methods is shown in table IV:

TABLE IV. ACCURACY OF DIFFERENT MODELS

modle	COAE2014 data set
SVM bow1 [12]	89.36%
SVM bow2 [12]	91.74%
SVM bow3 [12]	92.13%
CNN-word-static [12]	93.80%
CNN-Attentive(proposed)	95.15%

B. Experimental analysis

As shown in Table 4, the accuracy of SVM bow1,SVM bow2,SVM bow3 is improved sequentially. However, in SVM N-Gram models, the curse of dimensionality could appear in the case of N [12]. This causes the difficulty of the model training. The information of different text N-Gram can be extracted using convolution kernels with different sizes and the accuracy of the model is improved. The proposed network outperforms the above models. The deep level emotional semantic information obtained by attention probability distribution is coded so that the network performs better.

IV. CONCLUSIONS

Based on the CNN-Attention neural network, the context information of different levels is utilized effectively by the network. The curse of dimensionality is avoided in the proposed method. Text semantic information and rich text features are extracted using bottom layer. At the same time, the deep level emotional semantic information obtained by attention probability distribution is coded. In addition, effective corpus feature is extracted with less layers and the loss of information is reduced in feature extraction. Experimental results show improved accuracy in COAE2014 task 4 micro-blog data set for emotional classification up to 95.15%.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No.2014BAK11B02 and 2016 Guangxi Science and Technology Project, No.AB16380264.

REFERENCES

[1] Kim, Y. (2014). Convolutional neural networks for sentence classification. Eprint Arxiv.

[2] Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. Eprint Arxiv, 1.

[3] Johnson, R., & Zhang, T. (2014). Effective use of word order for text categorization with convolutional neural networks. Eprint Arxiv.

[4] T Luong， R Socher， CD Manning. Better Word Representations with Recursive Neural Networks for Morphology. Proceedings of the CoNLL-2013,2013,104.

[5] Zheng, X., Chen, H., & Xu, T. (2013). Deep learning for Chinese word segmentation and POS tagging. Conference on Empirical Methods in Natural Language Processing. [6] Socher, R., Bauer, J., Manning, C. D., & Ng, A. Y. (2013). Parsing with Compositional Vector Grammars. Meeting of the Association for Computational Linguistics (pp.455-465).

[7] Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. Computer Science, 4, 1188-1196.

[8] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.

[9] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[10] Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. Computer Science.

[11] Santos, C. D., Tan, M., Xiang, B., & Zhou, B. (2016). Attentive pooling networks.

[12] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26, 3111-3119.

[13] LIU Longfei,YANG Liang,ZHANG Shaowu,LIN Hongfei ， Convolution Neural Networks for Chinese Micro-blog Sentiment Analysis， journal of Chinese information processing ,Vol. 29, No. 6, November 2015