DOCUMENT CLASSIFICATION BASED ON CONVOLUTIONAL NEURAL NETWORK AND HIERARCHICAL ATTENTION NETWORK

Y. Cheng[*], *Z. Ye*[†], *M. Wang*[†], *Q. Zhang*[†]

Abstract: Numerous studies have demonstrated that the neural network model can achieve satisfactory performance in various natural language processing (NLP) tasks. In recent years, document classification is one of the NLP tasks that has gain considerable attention from researchers. For NLP tasks, convolutional neural network (CNN), recurrent neural network (RNN) and attention mechanism can be used. In this work, it is assumed that a document can be divided into two levels, word level and sentence level. In this paper, an effective and novel model called C-HAN (Convolutional Neural Network-based and Hierarchical Attention Network with RNN as basic units-based model) is proposed for document classification by combining the advantages of CNN, RNN and attention model. The CNN is used to extract the abstract relations between different words that are then fed into an attention based bidirectional long short-term memory recurrent neural network (Bi-LSTM) to obtain the high-level abstract representation of sentences. The representation of a document consists of sentences is obtained by using another attention based Bi-LSTM. Lastly, the classification ability of the proposed C-HAN model is evaluated on two datasets. The experimental results demonstrate that the C-HAN model outperforms previous deep learning methods and achieves the state-of-art performance.

Key words: *convolutional neural network, bidirectional long short-term memory, attention mechanism, document classification*

Received: 2018-08-27 **DOI:** 10.14311/NNW.2019.29.007
Revised and accepted: 2019-04-30

1. Introduction

In natural language processing, text classification has always been a fundamental task aiming at inferring the label of a given text. Text classification has broad applications including sentiment classification [1], spam detection [2], question classification [3], and news topic classification [4]. Knowledge-based approaches has been widely used in NLP tasks, but these approaches require complex rules written manually to enable the computers understand human languages precisely. Due

[*]Yan Cheng – corresponding author, School of Computer Information Engineering, Jiangxi Normal University, NanChang, China chyan88888@jxnu.edu.cn,
[†]Z. Ye, M. Wang, Q. Zhang

to the complexity of human languages, knowledge-based approaches could only achieve certain results on small scale data [5]. Since 90s, machine learning based on statistics has begun to emerge in sentiment analysis [6,7]. However, these methods belong to "shallow learning". The function model or the calculation method is relatively simple but cannot express some complex functions. Overall, the machine learning methods always have poor generalization performance and require artificial selection of a large amount of features of data. These defects cause the machine learning methods to hit bottlenecks in NLP tasks. Deep learning can automatically select the appropriate features and expressions from the original data to process a variety of complex tasks. These methods have obvious advantages in modeling, interpretation and optimization. The traditional text classification methods always represent the sentences or documents with n-gram features and use a simple linear model to realize the function of text function [8]. In recent years, various approaches based deep learning methods, such as convolutional neural networks (CNN) [9], recurrent neural networks (RNN), long short-term memory (LSTM) [10] and gated recurrent unit (GRU) [11] has been proposed to learn text representations with the continuous breakthrough of deep learning. Currently, attention model has become a promising model in text classification as a new model. In recent years, the use of deep learning model to analyze texts has become an attractive research field.

CNNs have achieved remarkable performance in computer vision [9], natural language processing [12,13] and speech recognition [14]. CNNs have the ability to capture local and spatial correlations of different input sources such as different words. In a number of sentence modeling tasks, CNNs perform excellently in extracting the contextual links of different words using some convolutional filters and identify the most significant features of a sentence through pooling operation [12]. RNNs are another widely used neural network architecture. RNNs can capture long-term dependencies of documents over time. LSTM, as an efficient variant of RNNs, was designed to avoid the frequent occurring problem of gradient vanishing in machine learning tasks. As a sequence model, the RNN has achieved remarkable performance in modeling sentences or documents [15,16]. The latest development in deep learning is attention model [17,18] that can capture the most important features in texts and optimize the model structures.

Obviously, a document consists of sentences and sentences consists of words in like manner. For sentiment analysis, the sentiment orientation of a document is determined by two levels, sentence level and word level. A document is made up of some sentences and different sentences have different degrees of importance for the result of sentiment orientation analysis. If the overall sentiment orientation of the document is positive, some of the more negative sentences are not important. Thus, the final result is not affect by all sentences. Then similarly, sentences are made up of words and different words have different degrees of influence on the judgment of the sentiment orientation of sentences. The existing models rarely explore the sentiment of documents from this point of view, and fail to reflect the influence of the hierarchical structure of documents on the result of the sentiment analysis.

Therefore, this paper proposes a new model C-HAN, composed of CNN, RNN and attention models. The CNN is constructed to extract the n-grams features

among the pre-trained word vectors in sentences to learn the higher-level representations of n-grams features and the position of words. Then the obtained new word vectors after convolution operation serve as the input of Bi-LSTM. Next, the attention model is used to determine these features with abundant information. Similarly, for sentences, the vector representation of each sentence is obtained. The attention model is then again to determine the most informative sentences in the document. In conclusion, the proposed model can be trained end to end totally does not rely on any complicated pre-processing or external language knowledge.

2. Related Work

Deep learning methods provide excellent performance in various NLP tasks including sentiment classification [1], sentence and document representation [19], and machine translation [20], etc. Distributed sentence representations learned by neural network models require a small amount of external domain knowledge and can achieve satisfactory performance in these sentiment classification tasks. CNN and RNN are the two most widely used variants in text classification tasks too.

In 2011, Collobert et al. [21] first proposed that the CNN could be used in POS tagging and other NLP tasks. Later in 2014, Kim [12] proposed the application of CNN for sentiment analysis tasks, and Kalchbrenner et al. [22] proposed wide convolution operation and k-max pooling. Recently, Conneau et al. [13] proposed a VDCNN model using the deep convolution network in sentiment classification. The RNN can capture better long-distance dependency between words in sentences compared with CNN. In recent years, numerous RNN-based models have been proposed. Tang et al. [23] proposed a hierarchical RNN model to build the relationship between sentences. Wang et al. [24] presented a DRNN model, which incorporates position-invariance into RNN by limiting the distance of information flow. Since both the CNN and the RNN have their own advantages, several approaches combining both network structures have been applied to NLP tasks. Lai S et al. [25] constructed a RCNN model that first used Bi-RNN model to obtain the representation of context, and then performed convolution and pooling operations to produce classification results. Zhou C et al. [26] proposed the C-LSTM model. Firstly, the CNN was used to extract the features of texts, and then a LSTM layer and a softmax layer were used to obtain the classification results. Attention mechanism is able to capture the importance of features about words or sentences, and it is also widely used with CNN or RNN in many NLP or CV tasks. Yang et al. [27] proposed hierarchical attention model for sentiment analysis. In short, deep learning methods have certain advantages in innumerable tasks and can be applied to the document classification, eliminating the tedious feature engineering operations of traditional methods.

3. The Proposed C-HAN Model

3.1 Convolutional Neural Network

In recent years, CNN has been proved to achieve excellent performance in many text classification tasks. It is well-known that convolution can be used to extract features and a pooling operation is often a step after the convolution. However, pooling operation is not used after the convolution in this paper because a Bi-LSTM is used instead. If pooling is selected as the necessary operation, the original sequence organization will be broken. The structure of CNN is shown in the following Fig. 1.

The purpose of padding in the Fig. 1 is to ensure that the dimension of the sentence representation matrix is the same as that of the text vector matrix. In CNN architecture, suppose that the current input $\mathbf{x}_{ij} \in R^d$ was the j-th word of the i-th sentence, where d denotes the dimension of word vectors. In this paper, each word is assigned a position vector $\mathbf{z}_{ij} \in R^d$, which is independent of the semantics of the word vector. Since the attention model is introduced in the proposed model, the position vector $\mathbf{z}_{ij}$ is used to assign attention weights to words according to their positions. These position vectors are learned through model training and they can help to maintain the position information of the word vectors. In this way, each word has a new encoding as follows:

$$\mathbf{a}_{ij}=\mathbf{x}_{ij}+\mathbf{z}_{ij}$$

Let the convolution kernel $\mathbf{B}^n \in R^{h \cdot d}$ $(n \in \{1,2,\ldots,d\})$ convolute on the new word vector encoding matrix, in which h is the length of the convolution kernel. A particular vector would be available after the convolution kernel convoluted on a word vector encoding matrix, which is called a feature map. The width of the convolution kernel is set equal to the dimension of word vector and n denotes the number of convolution kernels. Assuming that each $\mathbf{a}_{ij}$ of the corresponding vector is represented as a particular feature map $\mathbf{l}_{ij}=\left[l_{ij}^1, l_{ij}^2, \ldots, l_{ij}^d\right] \in R^d$ by a convolution

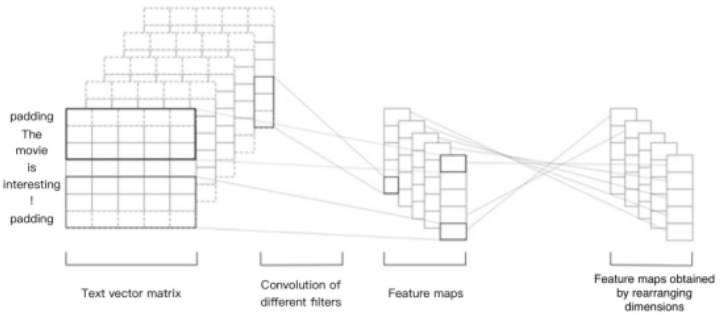

Fig. 1 *The architecture of convolutional neural network (CNN) model*

operation, the calculation procedure for each element in l_{ij} is as follows:

$$l_{ij}^{n} = f(\sum_{q=-(h-1)/2}^{(h-1)/2} \mathbf{a}_{ij+q} \bigodot \mathbf{B}_{(h-1)/2+q}^{n})$$

where $\bigodot$ is element-wise multiplication. In this paper, ReLU is selected as the nonlinear function and a valid way to generate feature maps. The model can use multiple filters (with same or varying window sizes) to obtain multiple features.

3.2 Bidirectional Long Short-Term Memory Networks

LSTM was originally designed to access the previous context over time so the future context is always ignored. However, it is meaningful for computers to distinguish the orientation of sentiment by considering the future context in text analysis. Therefore, the bidirectional LSTM is selected here. The Bi-LSTMs are able to read the texts from front to back and from behind to process the sequence data and then feed forward to the same output layer. In this paper, we use $\rightarrow$ and $\leftarrow$ denote the forward and backward processes, respectively. Let $\mathbf{h}^t$ be the hidden state that would be influenced by current data $\mathbf{x}^t$, the input gate $\mathbf{i}^t$, the forget gate $\mathbf{f}^t$, the output gate $\mathbf{o}^t$, the memory cell $\mathbf{c}^t$ and the previous hidden state $\mathbf{h}^{t-1}$ at each time step t. The following equations define the function of bidirectional LSTMs.

$$\overrightarrow{\mathbf{i}}^t = \sigma\left(\overrightarrow{\mathbf{W}}^i \overrightarrow{\mathbf{x}}^t + \overrightarrow{\mathbf{V}}^i \overrightarrow{\mathbf{h}}^{t-1} + \overrightarrow{\mathbf{b}}^i\right)$$

$$\overrightarrow{\mathbf{f}}^t = \sigma\left(\overrightarrow{\mathbf{W}}^f \overrightarrow{\mathbf{x}}^t + \overrightarrow{\mathbf{V}}^f \overrightarrow{\mathbf{h}}^{t-1} + \overrightarrow{\mathbf{b}}^f\right)$$

$$\overrightarrow{\mathbf{o}}^t = \sigma\left(\overrightarrow{\mathbf{W}}^o \overrightarrow{\mathbf{x}}^t + \overrightarrow{\mathbf{V}}^o \overrightarrow{\mathbf{h}}^{t-1} + \overrightarrow{\mathbf{b}}^o\right)$$

$$\overrightarrow{\mathbf{c}}^t = \overrightarrow{\mathbf{f}}^t \bigodot \overrightarrow{\mathbf{c}}^{t-1} + \overrightarrow{\mathbf{i}}^t \bigodot \tanh\left(\overrightarrow{\mathbf{W}}^c \overrightarrow{\mathbf{x}}^t + \overrightarrow{\mathbf{V}}^c \overrightarrow{\mathbf{h}}^{t-1} + \overrightarrow{\mathbf{b}}^c\right)$$

$$\overrightarrow{\mathbf{h}}^t = \overrightarrow{\mathbf{o}}^t \bigodot \tanh(\overrightarrow{\mathbf{c}}^t)$$

$$\overleftarrow{\mathbf{i}}^t = \sigma\left(\overleftarrow{\mathbf{W}}^i \overleftarrow{\mathbf{x}}^t + \overleftarrow{\mathbf{V}}^i \overleftarrow{\mathbf{h}}^{t+1} + \overleftarrow{\mathbf{b}}^i\right)$$

$$\overleftarrow{\mathbf{f}}^t = \sigma\left(\overleftarrow{\mathbf{W}}^f \overleftarrow{\mathbf{x}}^t + \overleftarrow{\mathbf{V}}^f \overleftarrow{\mathbf{h}}^{t+1} + \overleftarrow{\mathbf{b}}^f\right)$$

$$\overleftarrow{\mathbf{o}}^t = \sigma\left(\overleftarrow{\mathbf{W}}^o \overleftarrow{\mathbf{x}}^t + \overleftarrow{\mathbf{V}}^o \overleftarrow{\mathbf{h}}^{t+1} + \overleftarrow{\mathbf{b}}^o\right)$$

$$\overleftarrow{\mathbf{c}}^t = \overleftarrow{\mathbf{f}}^t \bigodot \overleftarrow{\mathbf{c}}^{t+1} + \overleftarrow{\mathbf{i}}^t \bigodot \tanh\left(\overleftarrow{\mathbf{W}}^c \overleftarrow{\mathbf{x}}^t + \overleftarrow{\mathbf{V}}^c \overleftarrow{\mathbf{h}}^{t+1} + \overleftarrow{\mathbf{b}}^c\right)$$

$$\overleftarrow{\mathbf{h}}^t = \overleftarrow{\mathbf{o}}^t \bigodot \tanh\left(\overleftarrow{\mathbf{c}}^t\right)$$

Then the concatenated vector of the outputs can be defined as follows:

$$\mathbf{h}^t = \overrightarrow{\mathbf{h}}^t \bigoplus \overleftarrow{\mathbf{h}}^t$$

Fig. 2 and Fig. 3 show the architectures of a unidirectional and a sequence of bidirectional LSTM models.

3.3 Attention Model

The attention model has been applied in machine translation tasks since 2014. The basic architecture of attention model is shown in Fig. 4.

As shown in the Fig. 4, Module1 is generally an encoder that performs a transition on the input data and Module2 is a decoder performing another transition on the data processed by Module1. Each output m_i is defined as:

$$m_i = F\left(C_i, m_1, m_2, \cdots m_{i-1}\right)$$

where C_i is the semantic encoding corresponding to each input data:

$$C_i = \sum_{j=1}^{T} a_{ij} S\left(n_j\right)$$

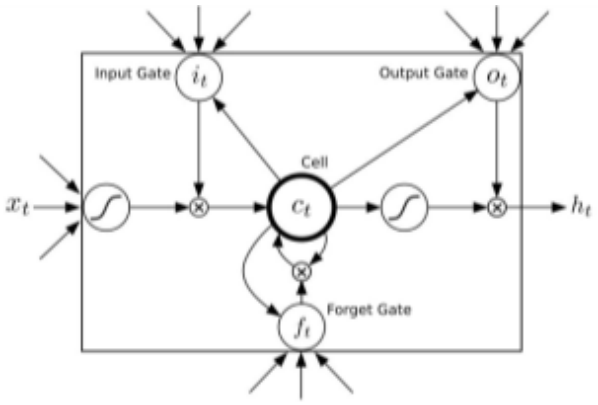

Fig. 2 *The architecture of a unidirectional LSTM model*

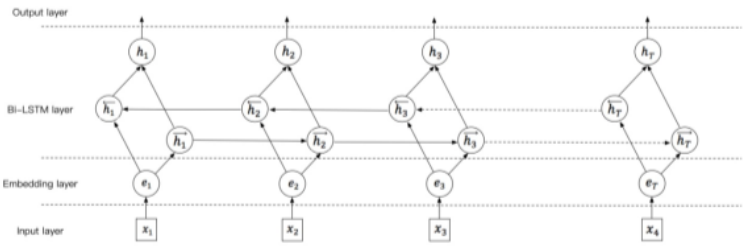

Fig. 3 *The architecture of a Bi-LSTM model*

Let $S(n_j)$ be the hidden status of the input data after it was processed by Module1, T is the number of input data and a_{ij} is the assigning probability of attention. The computation of a_{ij} is given by:

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})}$$

$$e_{ij} = \tanh(\mathbf{W}h_j + \mathbf{U}s_{i-1} + b)$$

Where e_{ij} refers to the influence evaluation score, h_j is the hidden state of input j in Module1, and s_{i-1} is the output of Module2 in the previous time step. $\mathbf{W}$ and $\mathbf{U}$ are the two weight matrices in the same way, and b is the offset value.

3.4 C-HAN Model

The architecture of the proposed C-HAN model is shown in Fig. 5. The proposed model consists of CNN and hierarchical attention network with Bi-LSTM as basic units.

The C-HAN model consists of four parts. First, the C-HAN converts the input words into the corresponding word vectors through the operation of the CNN layer. According to Section 3.1, the vector representation of the sentences after the convolution operation is obtained as $\mathbf{l}_{ik}$. After the first step is completed, the output of the CNN would be taken as the input of the Bi-LSTM network to obtain the output of the hidden layer, denoted as:

$$\overrightarrow{\mathbf{g}}_{ik} = \overrightarrow{LSTM}(\mathbf{l}_{ik})$$

$$\overleftarrow{\mathbf{g}}_{ik} = \overleftarrow{LSTM}(\mathbf{l}_{ik})$$

$$\mathbf{g}_{ik} = [\overrightarrow{\mathbf{g}}_{ik}, \overleftarrow{\mathbf{g}}_{ik}]$$

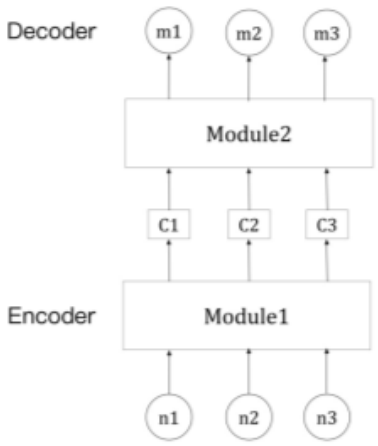

Fig. 4 *The architecture of attention model*

where the k value is between 1 and K, $\mathbf{g}_{ik}$ is a vectorized representation obtained after the module processing. As it is known that not all words are important to a sentence. The attention model is used to help to effectively determine the words in a sentence that contribute the most to the meaning of the sentence.

The attention model applied first to select the words that are meaningful to a sentence and to group their representations to form sentence vectors as:

$$\mathbf{u}_{ik} = \tanh\left(\mathbf{W}_w \mathbf{g}_{ik} + \mathbf{b}_w\right)$$

$$a_{ik} = \frac{\exp\left(\mathbf{u}_{ik}^T \mathbf{u}_w\right)}{\sum_k \exp(\mathbf{u}_{ik}^T \mathbf{u}_w)}$$

$$\mathbf{s}_i = \sum_k a_{ik} \mathbf{g}_{ik}$$

The word annotation $\mathbf{g}_{ik}$will be fed to obtain $\mathbf{u}_{ik}$ as a representation of $\mathbf{g}_{ik}$ through a one-layer MLP at first. Then, a normalized importance weight a_{ik} to weigh the importance of a word would be obtained using the softmax function by measuring the similarity of $\mathbf{u}_{ik}$ and a context vector $\mathbf{u}_w$. Next, the sentence vector s_i denoting the weighted sum of the word annotations is computed. It is worth noting that the word context vector $\mathbf{u}_w$ was randomly initialized and jointly learned with the training of the model. In a similar way, a document vector $\mathbf{d}_i$ can be obtained. A Bi-LSTM is used to encode the sentences in the proposed model:

$$\overrightarrow{\mathbf{d}}_i = \overrightarrow{LSTM}\left(\mathbf{d}_i\right)$$

$$\overleftarrow{\mathbf{d}}_i = \overleftarrow{LSTM}\left(\mathbf{d}_i\right)$$

$$\mathbf{d}_i = [\overrightarrow{\mathbf{d}}_i, \overleftarrow{\mathbf{d}}_i]$$

The attention model is utilized to measure the importance of sentences by introducing another sentence level context vector $\mathbf{u}_s$ as follows:

$$\mathbf{u}_i = \tanh\left(\mathbf{W}_s \mathbf{d}_i + \mathbf{b}_s\right)$$

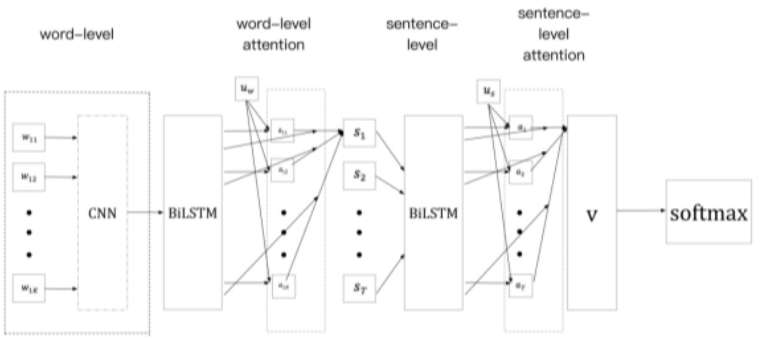

Fig. 5 *The architecture of the proposed C-HAN model for document classification*

$$a_i = \frac{\exp\left(\mathbf{u}_i^T \mathbf{u}_s\right)}{\sum_i \exp\left(\mathbf{u}_i^T \mathbf{u}_s\right)}$$

$$\mathbf{v} = \sum_i a_i \mathbf{d}_i$$

Thus, the final document vector $\mathbf{v}$ can be obtained. This is an abstract representation of a document and it can be used as the important feature for document classification:

$$p = \text{softmax}\left(\mathbf{W}_1 \mathbf{v} + \mathbf{b}_1\right)$$

For text classification, the cross-entropy function is used as cost function and to train the proposed model. Specifically, the negative log likelihood function is selected as training loss:

$$L = -\sum_z \log p_{zj}$$

where j denotes the label of a document z.

4. Experiments

4.1 Data sets

The proposed C-HAN model is evaluated on two data sets: Yelp reviews and IMDB reviews. Both data sets are briefly introduced below.

Yelp reviews: The data set is obtained from the Yelp Dataset Challenge in 2015 [23]. There are five levels of ratings from 1 to 5.

IMDB reviews: The data set is obtained from IMDB [28]. In contrast to above-mentioned data set, IMDB has ten levels of ratings from 1 to 10.

4.2 Model configuration and training

The proposed model is implemented based on Keras [29] which is a python library. In this paper, NLTK [30] is used to split the documents into sentences and tokenize each sentence into words. A word2vec [31] model is trained on the training and validation sets to obtain the word embedding to initialize the weights. The dimension of the word vector is set to be 300. The dimension of the position vector is set equal to that of the word vector and would be randomly initialized. While building a vocabulary, the words appearing more than five times are retained, and the words appearing less than five times are replaced with UNK tokens that are randomly initialized. Then the maximum sentence length is set to 50. Padding or truncated operations are performed if the length of the sentence is less than 50 or more than 50, respectively. Similarly, the maximum number of sentences in a document is set to 20, and the pre-processing is also done with zero padding and truncation.

The overall model architecture consists of a convolution layer, a bidirectional LSTM with attention layer in analysis of a sentence and another bidirectional LSTM with attention layer in analysis of a document. Other model architectures such as using two convolution layers are also tried. The "valid" convolution model

is adopted in which the number of filters is set to 300 and the size of convolution window is set to 3. The experiment shows that the model can achieve state-of-art performance with above-mentioned parameter configuration. For the bidirectional LSTM with attention layer, the dimension is set to 300 and the context vector is initialized randomly. For training, the mini-batch size is set to 32 and gradient update strategy called stochastic gradient descent is used. The optimization method used in the proposed model is Adam [32] and the initialization rate is set to 0.001. In addition, the dropout [33] is used to avoid the over-fitting problem with the masking probability is set as 0.5.

4.3 Results and Model Analysis

In this section, the performance assessment of the proposed model is presented for the two data sets. In addition, the analysis of the proposed model regarding the selection of the filters and the analysis of attention layers are also presented.

The baseline models generally include machine learning models, convolution neural network model, LSTM model, and HAN. In this paper, SVM [12] is selected in machine learning models due to its efficiency in many tasks. Furthermore, the character-level CNN, the word-level CNN [12]. In addition, we choose LSTM, Conv-GRNN, LSTM-GRNN and HAN [27] are selected as baseline models for comparison with the proposed model. The experimental results on the two data sets are shown in Tab. I:

It can be seen from Tab. I that the basic deep learning models, such as CNN-char, CNN-word, LSTM and Bi-LSTM have little advantage over the traditional methods like SVM for text classification tasks because they cannot properly utilized the hierarchical document structure or attention mechanism. However, the traditional models like SVM and others often highly depend on engineered features. These features not only require manual labor, but also cause error propagation due to some defect in the tools. Therefore, the traditional models cannot be well extended to other data sets or other tasks. However, the basic deep learning methods have the ability to automatically learn the semantic sentence representation with-

Models	Yelp	IMDB
SVM + Unigrams	61.1	39.9
SVM + Bigrams	62.4	40.9
CNN-char	62.0	39.3
CNN-word	60.5	37.6
LSTM	58.2	35.5
Bi-LSTM	58.0	36.2
Conv-GRNN	66.0	42.5
LSTM-GRNN	67.6	45.3
HAN	71.0	49.4
C-HAN(proposed model)	72.2	49.8

Tab. I *Comparison of the proposed model with the baseline models on Yelp 2015 and IMDB*

out any manual design features and can save a lot of time. It can also be seen that the CNN-char provides accuracy of 62.0 and 39.3 for Yelp 2015 and IMDB, respectively, while the Bi-LSTM has accuracy of 58.0 and 36.2 for Yelp 2015 and IMDB, respectively. Morevoer, it can be observed that combining these neural network based models will significantly enhance the accuracy. The Conv-GRNN outperforms the CNN-char by 5.5% and 3.2% than CNN-char on Yelp 2015 and IMDB, respectively. The LSTM-GRNN performs even more better. The HAN model that adds hierarchy and attention model to the traditional neural network, can achieve the best performance. For example, the HAN outperforms the LSTM-GRNN by 3.4% and 4.1% on Yelp 2015 and IMDB, respectively. These results indicate that the hierarchical split of documents is effective. The proposed model first uses a CNN layer to extract the higher-level features among words as input to the next module. Thus, on the one hand, the C-HAN has the advantages of combining the neural networks, and on the other hand, it also embodies the advantages of hierarchy. The result demonstrate that the proposed C-HAN model provides accuracy of 72.2 and 49.8 for Yelp 2015 and IMDB, respectively, while the HAN has accuracy of 71.0 and 49.4 fro Yelp 2015 and IMDB, respectively. Specially, the C-HAN outperforms the HAN by 1.2% and 0.4% than HAN on Yelp 2015 and IMDB, respectively.

4.4 Filter Selection and Attention Mechanism Analysis

This section will analyze the influence of different selection of filters on the proposed C-HAN model performance and will demonstrate the effectiveness of the attention mechanism. First, the convolution layer is used to extract n-gram features of texts using several different filter sizes.

Tab. II shows the prediction accuracies on the Yelp 2015 data set. The results show that the proposed model provides perform best when single convolution layer with filter length 3 is used.

It can be seen from Tab. I that the accuracy declines and drops below 70% when filter length with different sizes like 2, 3 and 4 is used. Therefore, it is important that single convolution layer with filter length 3 is used to capture the tri-gram features from the two data sets used in this paper. It is believed that this result can be a reference for other tasks in NLP using deep learning methods.

As it is necessary to visualize the weight of attention to embody the effect of a

Filter configurations	Yelp
2	70.6
3	72.2
4	70.1
2,3	68.4
2,4	68.6
3,4	68.8
2,3,4	66.9

Tab. II *Prediction accuracies on Yelp 2015 using different filter sizes*

new model using attention mechanism. Therefore, the attention weight is visually presented. For example, four texts from IMDB datasets and Yelp 2015 data sets are presented below.

Example 1: 'hollywood hotel was the last movie musical that busby berkeley directed for warner bros his directing style had changed or evolved to the point that this film does not contain his signature overhead shots or huge production numbers with thousands of extras', 'by the last few years of the thirties swing style big bands were recording the years biggest popular hits', 'the swing era also called the big band era has been dated from 1935 to 1944 or 1939 to 1949 although it is impossible to exactly the moment that the swing era began benny goodmans engagement at the ballroom in los angeles in the late summer of 1935 was certainly one of the early that swing was entering the consciousness of mainstream americas youth', 'when goodman featured his swing repertoire rather than the society style dance music that his band had been playing the youth in the audience went wild', 'that was the beginning but since radio live concerts and word of mouth were the primary methods available to spread the phenomena it took some time before swing made enough to produce big hits that showed up on the pop charts', 'in hollywood hotel the appearance of benny goodman and his orchestra and raymond paige and his orchestra in the film indicates that the film industry was ready to capitalize on the shift in musical taste the film was in production only a year and a half or so after goodmans ballroom engagement', 'there are a few interesting musical moments here and there in hollywood hotel but except for benny goodman and his sing sing sing there isnt a lot to commend', 'otherwise the most interesting musical sequences are the opening hooray for hollywood parade and let that be a lesson to you production number at the drive in restaurant', 'the film is most interesting to see and hear benny goodman and his orchestra play and dick powell and frances sing'.

Example 2: 'as you may or may not have heard there is no actual fighting between vampires or zombies in this film', 'one may then ask why the title suggested such a thing but really its kind of appropriate because nothing else about this film made any sense either', 'one may then ask why the title suggested such a thing but really its kind of appropriate because nothing else about this film made any sense either', 'the acting was incredibly bad worse than commercials bad not only were there no fighting between vampires and zombies but i think there was only one scene with zombies even in it', 'their make up looked as if it were applied by an 8 year old girl', 'the scene was totally random and out of place and featured one of the characters fighting the zombies off with a im not kidding but they used chainsaw sound effects this was undoubtedly the poorest movie ive ever seen in my life', 'the only circumstance that i wouldnt totally ridicule every person responsible for production of this film is if i learned that it was produced entirely by 11 year olds really though even with all of the criticism i offer here id suggest watching this movie solely based on the fact that it may very well be the worst movie ever and because of this is quite comical', 'even just counting the flaws in it should keep you entertained'.

Example 3: 'this should be a great film meryl streep and jack nicholson co starring as two newspaper writers', 'mike directing', 'uh uh', 'its dull dull dull', 'pointless and predictable', 'slow and unfocused its a cookie cutter boy meets girl

boy marries girl boy has affair girl leaves boy story', 'now theres an original concept', 'after through two hours was it only two', 'it felt like six', 'i wasnt sure whether it was a comedy a romance a tragedy or a soap opera', 'it was done in 1986 im sure all of us did things sixteen years ago that we rather would forget', 'i hope the damage to the reputations of streep et al is beginning to heal and that the on the master is beginning to fade', 'its not that its such a bad picture', 'its just that its such an un good one'.

Example 4: 'walter matthau and george burns just work so well together', 'the of willy with the perplexed al is a mixture made in heaven', 'the scene when they meet again in flat is a gem and the final scene rounds up the film to perfection', 'walter matthau gives a superb performance as the irascible semi retired comedian as only he can the in the voice and the exaggerated dramatics coupled with his general misunderstanding of what is going on form a great characterization', 'george burns timing is legendary and nowhere was it better than in this film his calm aplomb with desert dry replies are memorable', 'watch for the scene near the end when al and his daughter ask something of the spanish caretaker and als reaction priceless'.

Fig. 6 shows the attention weight maps for each of the four examples presented above.

In Example 1, there are nine sentences and it can be seen that the ninth sentence is more important for the judgement of the sentiment orientation analysis. Referring to the original text of Example 1, the words 'most' and 'interesting' can be seen in the last sentence. It shows that the proposed model does find important information. When a sentence is negative, the proposed model is also applicable. By analyzing Example 2, it can be seen that the penultimate sentence is crucial to the sentiment of texts and facts also prove it. Similarly, the remaining two examples can also be analyzed.

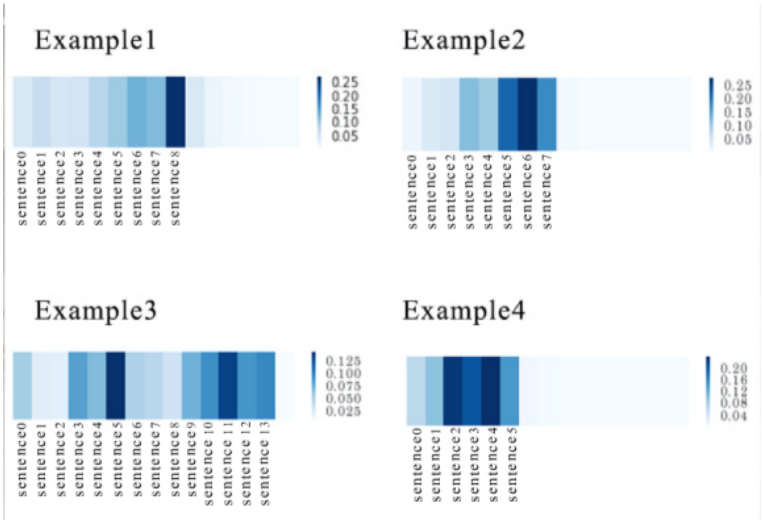

Fig. 6 *The attention weight maps of four example texts*

5. Conclusion

In this paper, a new hierarchical model called C-HAN is proposed. The C-HAN combines the convolutional neural networks with attention-based bidirectional LSTM. The proposed model learns n-grams features using the convolution layer and then feeds them to the attention-based hierarchical model. At the same time, using hierarchical attention network with Bi-LSTM as basic unit, the proposed model is evaluated using Yelp 2015 and IMDB data sets and compared with baseline models. The experimental results validate that the proposed model can achieve superior performance compared to existing models. Future work can include exploration of other forms of attention mechanisms based on this work.

Acknowledgement

This research has been supported by the National Natural Science Foundation of China (grant No. 61262080 and No. 61562043), the Key Science and Technology Foundation of Jiangxi province (grant No. 20151BBE50121 and No. 20161BBE50086) and the Foundation of Jiangxi Educational Committee (grant No. GJJ150299).

References

[1] PANG B., LILLIAN L. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*. 2(1-2), pp. 1-135, 2008, doi: 10.1561/1500000011.

[2] SAHAMI M., DUMAIS S., HECKERMAN D., et al. A bayesian approach to filtering junk e-mail. In: *Learning for Text Categorization: Papers from the 1998 workshop*. 62, pp. 98-105, 1998.

[3] MOSCHITTI A., QUARTERONI S., BASILI R., et al. Exploiting syntactic and shallow semantic kernels for question answer classification. In: *Proceedings of the 45th annual meeting of the association of computational linguistics*. pp. 776-783, 2007.

[4] RICHARD S., TORU I., FRANCIS K., et al. A maximum likelihood model for topic classification of broadcast news. In: *Proceedings of the European Conference on Speech Communication and Technology*. pp. 1455-1458, 1997.

[5] CLORE G.L., COLLINS A. The cognitive structure of emotions. *Contemporary Sociology*, 18 (6), pp. 2147-2153, 1988, doi: 10.2307/2074241.

[6] PANG B., LILLIAN L., VAITHYANATHAN S. Thumbsup? sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*. 10, pp. 79-86, 2002, doi: 10.3115/1118693.1118704.

[7] HU M., LIU B. Mining and summarizing customer reviews. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 168-177, 2004, doi: 10.1145/1014052.1014073.

[8] WANG S., CHRISTOPHER D., MANNING. Baselines and bigrams: simple, good sentiment and topic classification. In: *Proceedings of the Association for Computational Linguistics*. 2, pp. 90-94, 2012.

[9] KRIZHEVSKY A., SUTSKEVER I., HINTON G.E. ImageNet classification with deep convolutional neural networks. In: *Proceedings of International Conference on Neural Information Processing Systems*. 25, pp. 1097-1105, 2012, doi: 10.1145/3065386.

[10] HOCHREITER S., SCHMIDHUBER J. Long short-term memory. *Neural computation*. 9 (8), pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[11] CHO K., VAN M.B., GULCEHRE C., et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv*: 1406.1078, 2014, doi: 10.1074/jbc.M608066200.

[12] KIM Y. Convolutional neural networks for sentence classification. *arXiv preprint arXiv*:1408.5882, 2014.

[13] CONNEAU A., SCHWENK H., BARRAULT L., et al. Very deep convolutional networks for text classification. *arXiv preprint arXiv*:1606.01781, 2016.

[14] ABDEL-HAMID O., DENG L., YU D. Exploring convolutional neural network structures and optimization techniques for speech recognition. In: *Proceedings of Interspeech*. 2013, pp. 1173-5, 2013.

[15] MAO J., XU W., YANG Y., et al. Deep captioning with multimodal recurrent neural networks. *arXiv preprint arXiv*: 1412.6632, 2014.

[16] YIN W., KANN K., YU M., et al. Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv*: 1702.01923, 2017.

[17] BAHDANAU D., CHO K., Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv*: 1409.0473, 2014.

[18] YIN W., SCHUTZE H., XIANG B., et al. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *arXiv preprint arXiv*: 1512.05193, 2015.

[19] GEHRING J., AULI M., GRANGIER D., et al. Convolutional sequence to sequence learning. *arXiv preprint arXiv*: 1705.03122, 2017.

[20] KLEIN G., KIM Y., DENG Y., et al. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv*: 1701.02810, 2017.

[21] COLLOBERT R., WESTON J., BOTTOU L., et al. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*. 12 (8), pp. 2493-2537, 2011, doi: 10.1016/j.chemolab.2011.03.009.

[22] KALCHBRENNER N., GREFENSTETTE E., BLUNSOM P. A convolutional neural network for modelling sentences. *arXiv preprint arXiv*: 1404.2188, 2014.

[23] TANG D., QIN B., LIU T. Document modeling with gated recurrent neural network for sentiment classification. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. pp. 1422-1432, 2015, doi: 10.18653/v1/d15-1167.

[24] WANG B. Disconnected Recurrent Neural Networks for Text Categorization. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 1, pp. 2311-2320, 2018.

[25] LAI S., XU L., LIU K., et al. Recurrent Convolutional Neural Networks for Text Classification. In: *Proceedings of Twenty-ninth AAAI conference on artificial intelligence*. pp. 2267-2273, 2015, doi: 10.1145/2808719.2808746.

[26] ZHOU C., SUN C., LIU Z., et al. A C-LSTM Neural Network for Text Classification. *Computer Science*. 1 (4), pp. 39-44, 2015.

[27] YANG Z., YANG D., DYER C., et al. Hierarchical Attention Networks for Document Classification. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp. 1480-1489, 2016, doi: 10.18653/v1/n16-1174.

[28] DIAO Q., QIU M., WU C., et al. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 193-202, 2014, doi: 10.1145/2623330.2623758.

[29] CHOLLET F. Keras: The python deep learning library. *Astrophysics Source Code Library*, 2018, doi: 10.1086/316861.

[30] BIRD S., LOPER E. NLTK: the natural language toolkit. In: *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. pp. 31, 2004, doi: 10.3115/1225403.1225421.

[31] MIKOLOV T., SUTSKEVER I., CHEN K., et al. Distributed representations of words and phrases and their compositionality. In: *Proceedings of Advances in neural information processing systems*. pp. 3111-3119, 2013, doi: 10.1162/jmlr.2003.3.4-5.951.

[32] KINGMA D.P., BA J. A. A method for stochastic optimization. *arXiv preprint arXiv*: 1412.6980, 2014.

[33] SRIVASTAVA N., HINTON G., KRIZHEVSKY A., et al. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*. 15 (1), pp. 1929-1958, 2014.