IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 28, 2019, accepted March 9, 2019, date of publication March 20, 2019, date of current version April 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2906398

A Novel Capsule Based Hybrid Neural Network for Sentiment Classification

YONGPING DU, XIAOZHENG ZHAO, MENG HE, AND WENYANG GUO
College of Computer Science, Beijing University of Technology, Beijing 100124, China

Corresponding author: Xiaozheng Zhao (zxz53000@163.com)

This work was supported by the National Key Research and Development Program of China under Grant 2018YFC1900800, and in part by the Research Program of State Language Commission under Grant YB135-89.

ABSTRACT Sentiment classification of short text is a challenging task because of limited contextual information. We propose a capsule-based hybrid neural network model which can obtain the implicit semantic information effectively. Bidirectional gated recurrent unit (BGRU) is applied in this model to achieve the interdependent features with long distance. Moreover, the capsule network can extract richer textual information to improve expression ability. Compared with the attention-based model which combines self-attention mechanisms and convolutional neural networks (CNN), the capsule-based hybrid model has the advantage of less training time and simple network structure to achieve better performance. The performance is evaluated on two short text review datasets. Our capsule-based model outperforms other related models on movie review data and gets the highest accuracy of 0.8255. Meanwhile, it performs better than most of the systems in NLPCC2014 Task II and, especially achieves the best result on negative data.

INDEX TERMS Sentiment classification, capsule network, bidirectional gated recurrent unit, deep learning.

I. INTRODUCTION

Sentiment classification is an important task in Natural Language Processing (NLP) field. It is useful to understand user opinions in social networks or product reviews [1]. The task aims to determine the sentiment polarity of textual expressions, positive, negative or neutral. Unlike traditional expression, short text reviews have a high degree of colloquialism, non-standard grammatical structures and low sentence integrity. Therefore, it is more challenging to judge the sentiment polarity accurately.

There are many works which have achieved good results in the field of sentiment analysis. Though these previous methods have achieved good performance, sentence-level sentiment classification still remains a huge challenge for several reasons. Most machine learning methods overly rely on handcrafted features which require lots of manual design and adjustment, and it is time-consuming and cost-intensive. Though the problem is helped greatly by the proposal of deep learning in recent years, these neural network based approaches cannot encode and learn the part-whole relationship in the short text efficiently.

In order to address the above limitations, this paper presents a capsule-based hybrid neural network model for sentiment classification. The goal is to encode the intrinsic part-whole relationship and explore the grammatical and syntactic features to enrich the representation comprehensively. Furthermore, we also design an attention-based model for comparison, which leverage self-attention and CNN (Convolutional Neural Network) to learn contextual and local feature separately. The contributions of the paper are listed as follows:

- ✓ The semantic of each sentence is represented by BGRU (Bidirectional Gated Recurrent Unit), which shortens the distance between interdependent features.
- ✓ We propose a self-attention mechanism to obtain the dependencies between words directly and capture the internal structure of the sentence, which is combined with CNN to extract bigram and trigram features further.
- ✓ The capsule network with dynamic routing is devised to extract richer text information. It improves the text expression ability and acquires more important clues to improve sentiment classification performance.

II. RELATED WORK

The traditional sentiment classification methods mainly focus on sentiment dictionary or machine learning approach. Based

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

on massive micro-blog data, Zhao *et al.* [2] proposed a method to construct a large-scale sentiment dictionary and the experimental results showed that it was helpful to improve the sentiment classification performance. In addition, machine learning approach is used widely for sentiment classification task. Based on dependency parsing features and compositional semantic information, Odbal and Wang [3] proposed a phrase level emotion detection model based on semi-CRFs (semi-Markov Conditional Random Fields), which played an important role in implicit emotion detection. Li *et al.* [4] made use of various features and resources in an SVM classifier, and it also combined with probability-output weighting to improve the classification accuracy.

With greater attention being placed on deep learning in natural language processing, neural network models have been introduced in sentiment classification task due to their ability to carry out text representation learning. Deep learning method can achieve the potential semantic features on large-scale corpus. Convolution and Pooling of CNN (Convolutional Neural Network) can be well applied to extract local features [5]–[7]. Deriu *et al.* [8] leveraged large amounts of data with distant supervision to train the convolutional neural networks, whose predictions are combined with a random forest classifier, which optimizes the polarity classification. LSTM (Long Short-Term Memory) neural network can capture long-term dependencies in a sequence by introducing storage units and gate mechanisms, which decide how to use and update the information in the storage unit to obtain more permanent memory, so that it can increase the advantages of depth calculation [9]-[10]. Du *et al.* [11] combined LSTM model with target information, which improved the accuracy of target-dependent sentiment analysis significantly. Tai *et al.* [12] put forward a tree-structured long short-term memory networks to improve the semantic representations. Nowadays, the deep neural network model combined with attention mechanism has achieved better results than traditional methods in target-based tasks, such as classification of relationships based on specific evaluation objects [13], sentence pair modeling based on specific target [14], and machine translation based on specific target [15]. For sentiment analysis task, attention mechanisms are mainly applied on aspect-based sentiment classification [16]. Galassi *et al.* [17] proposed a new neural architecture that exploits readily available sentiment lexicon resources. Two kinds of attention, including lexicon-driven contextual attention and contrastive co-attention, are adopted to improve model performance.

Recently, Sabour *et al.* [18] proposed a capsule network for image classification. A capsule is a group of neurons whose activity vector represents the instantiation parameters of a specific entity type such as an object or an object part. An iterative dynamic routing algorithm is adopted. A lower-level capsule prefers to send its output to higher level capsules whose activity vectors have a big scalar product with the prediction coming from the lower-level capsule. Zhao *et al.* [19] proposed three strategies to stabilize the dynamic routing process to alleviate the disturbance of some noise capsules which may contain redundant information or have not been successfully trained. Kim *et al.* [20] proposed a simple routing method that effectively reduces the computational complexity of dynamic routing, which has achieved good results on multiple data sets. Zhang *et al.* [21] explored the capsule network with attention mechanism for relation extraction in a multi-label learning framework, and the performance is improved. Zhang *et al.* [22] introduced a Capsule network for sentiment analysis in domain adaptation scenario with semantic rules, which can enhance the comprehensive sentence representation learning. Such models not only pay special attention to the characteristic information during training process, but also adjust the parameters of the neural networks for different features effectively and mine more hidden characteristic information.

III. CAPSULE BASED HYBRID MODEL FOR SENTIMENT CLASSIFICATION

We propose a Capsule based Hybrid Model for sentiment classification and the structure is shown in Fig.1. It consists of five modules: Semantic Representation Module, Word Attention Module, Capsule Module, Feature Extraction Module and Classification Module. Assuming that the input sentence is $Z = [w_1, ...w_i, ...w_n]$, the goal of this model is to predict the sentiment polarity of sentence Z, which will be Positive(P) or Negative(N).

The Capsule based Hybrid Model is designed for sentiment classification task. Meanwhile, we also make use of self-attention mechanism giving more clues to capture important features by CNN for comparison. More importantly, the capsule based model fuses the role of self-attention and CNN simultaneously, which has the advantage of less training time and simple network structure to achieve a better performance than the comparison model.

A. SEMANTIC REPRESENTATION MODULE

Word Embedding: Represent each word as a multi-dimensional distributed vector. The sentence with n words is input into the embedding layer and each word is transformed to d-dimensional word vector. Ultimately, the embedding layer encodes the sentence representation as a matrix $Z = [w_1, ..., w_i, ..., w_n] \in R^{n*d}$, where $w_i = [x_{i1}, ..., x_{ij}, ..., x_{id}]$ corresponds to the word vector of the word $\mathbf{w_i}$ in the sentence.

The Long Short-Term Memory (LSTM) Neural Network was first proposed by Hochreiter and Schmidhuber [23]. It's a special Recurrent Neural Networks (RNN) and it has the ability to learn long-term dependencies, especially in text processing. It can not only predict the probability of next word in a language model by the contextual information [24], but also solve the problem of gradient disappearance in traditional RNN for long sequence data [25]. GRU (Gated Recurrent Unit) is a variant of LSTM. It simplifies the gating mechanism and speeds up the training. We adopt Bidirectional GRU to get the hidden representation of text quickly.

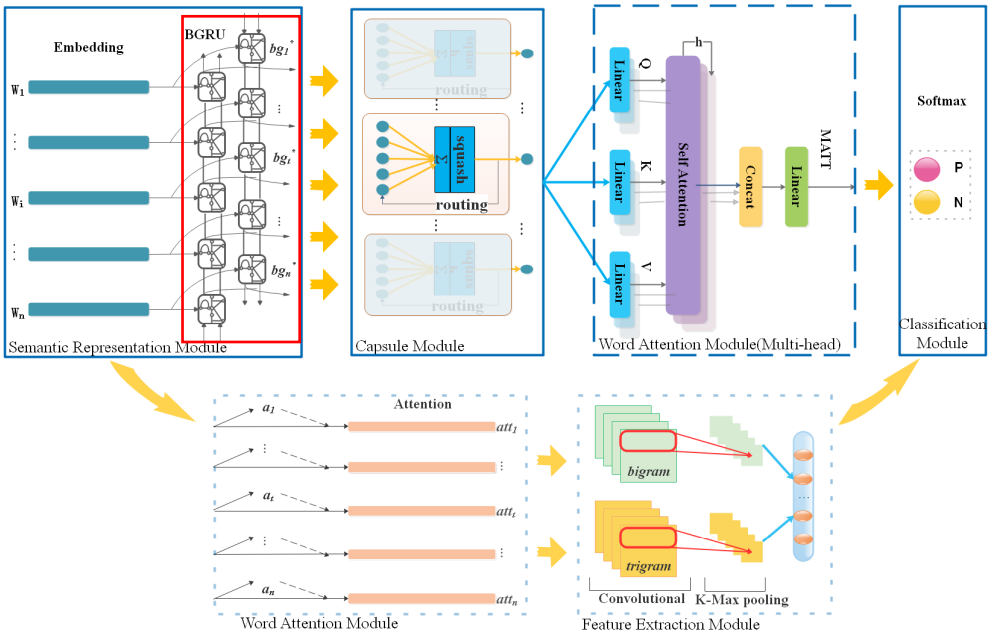


FIGURE 1. Capsule based Hybrid Model for Sentiment Classification.

B. WORD ATTENTION MODULE

Self-Attention can capture syntactic or semantic features between words in the same sentence effectively. Especially, any two words in a sentence will be linked directly and so it will be easier to obtain the interdependent features with long-distance.

In order to calculate the self-attention weight of each word, one layer perceptual network is used to calculate the output of BGRU bg_t^* at each time t. And the corresponding attention score g_t is calculated as Equation (1).

$$g_t = \tanh\left(W * bg_t^* + b\right) \tag{1}$$

where W and b indicate the weight matrix and bias of the perceptual network respectively.

Furthermore, the attention score at each time t is normalized by Equation (2).

$$a_t = \frac{\exp\left(g_t\right)}{\sum_{j=1}^{n} \exp\left(g_j\right)} \tag{2}$$

Each normalized attention score a_t is corresponding to BGRU output bg_t^* separately, and the final output of the self-attention model is the matrix $ATT = \{att_1, ..., att_t, ..., att_n\}$, where $att_t = bg_t^* * a_t$ and $a_t \in R^{d+d}$.

Vaswani *et al.*[26] proposes an attention mechanism called Multi-head Attention. Instead of performing a single attention function with d-dimensional Keys, Values and Queries, we find it is beneficial to linearly project the Keys, Values and Queries h times with different learned linear projections to d_k, d_v and d_q dimensions respectively. On each of these projected versions, the attention function is performed in parallel, yielding d_v-dimensional output value. They are concatenated and projected further, resulting in the final value. In this paper, we aim at sentiment classification in sentence-level, and so the input Keys, Values and Queries are the same as the output of each capsule v_i^*, which will be introduced in section 3.3. The output of the multi-head attention model is achieved by Equation (3).

$$MultiHead(v_i^*, v_i^*, v_i^*) = Concat(head_1, ..., head_h) \tag{3}$$

Here, each $head_i$ is also calculated by perceptual network and softmax function and the normalized attention score ma_i is got. Finally, the output matrix of the multi-head attention model is $MATT = \{matt_1, ..., matt_i, ..., matt_h\}$ where $matt_i = h * v_i^* * ma_i$ and $ma_i \in R^{d+d}$.

C. CAPSULE MODULE

Capsule Network is proposed by Hinton *et al.*[18] and [27]. The individual neuron node in traditional neural network is replaced by the neuron vector in capsule network which is trained by dynamic routing algorithm. We combine the capsule network and BGRU model to implement the sentiment classification. Capsule network can extract richer text information, and also the word position, semantic and syntactic structure can be encoded effectively. It improves the text expression ability and acquires more important clues further.

The input of capsule network is bg_t^* that is the output from previous layer BGRU. The operations are shown in Equation (4)-(6).

$$\hat{u}_{o|i} = W_{io} bg_t^* \tag{4}$$

$$s_{out} = \sum_{i=1}^{m} c_{io} \hat{u}_{o|i} \tag{5}$$

左边加入target

$$c_{io} = \frac{\exp(b_{io})}{\sum_k \exp(b_{ik})} \tag{6}$$

Here, W_{io} represents the weight matrix which controls the connection strength between the input and output layer. c_{io} is the coupling coefficient which is updated iteratively by dynamic routing algorithm. The sum of coupling coefficient between the capsules of input layer and output layer is 1. It is computed by softmax with initialized b_{io} to 0.

The non-linear activation function *squash* is adopted to normalize the output vector shown in Equation (7).

$$v_{out} = \frac{\|s_{out}\|^2}{1 + \|s_{out}\|^2} \frac{s_{out}}{\|s_{out}\|} \tag{7}$$

The process of dynamic routing algorithm is shown in Fig.2.

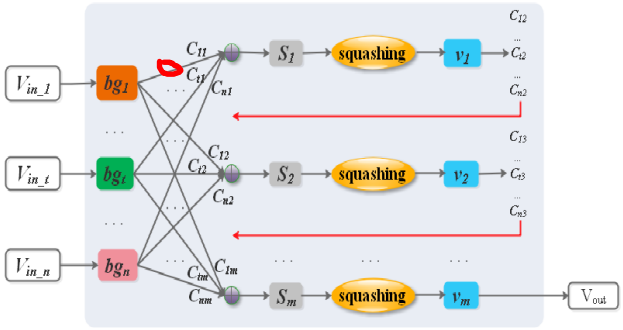


FIGURE 2. Dynamic Routing Process in Capsule Network.

The coupling coefficient vector is initialized as $c = [c_{11}, ..., c_{t1}, ..., c_{n1}]$ and the output v_1 is got according to equation (4)-(7). The vector c is updated iteratively by evaluating the influence of component bg_t on output v_i in each iteration. With the variation of vector c, the weight of important feature in the serialized text is increased. The larger of the coupling coefficient c_{io}, the higher weight value of important semantic feature, and it will give great help to correct classification.

D. FEATURE EXTRACTION MODULE

Convolutional Neural Network (CNN) is a special kind of deep neural network model. We design a double-layer parallel convolutional neural network to extract and represent the short text features.

1) CONVOLUTION LAYER FOR FEATURE EXTRACTION

The purpose of convolution layer is to extract semantic features of the sentence, each convolution kernel corresponds to a certain part of feature and the feature mappings can be obtained after convolution operation. The number of convolution kernels is set to 128 in our work. The convolution is operated on each sentence matrix $ATT = \{att_1, ..., att_t, ..., att_n\}$, which is the output of the previous word attention model, by Equation (8).

$$S = f(W * ATT + b) \tag{8}$$

where $f = relu = \max(0, x)$ is a nonlinear activation function, W and b represent the weight matrix and bias of the network respectively. S denotes the feature matrix extracted by the convolution operation.

The convolution window size is set to 2 and 3 respectively which can extract both the bigram and trigram features.

2) K-MAX POOLING LAYER FOR FEATURE DIMENSION REDUCTION

Features extracted by the convolution layer are transmitted to the pooling layer which will further aggregate and simplify the feature representation. K-Max pooling is adopted to select the top-K value of each filter to represent the semantic information. The larger the feature value, the greater the emotional strength. In addition, it will also preserve the relative order information of these features. The value of K is set to $\lfloor (len - f_s + 1)/2 \rfloor$, where len is the length of the sentence and f_s is the convolution window size.

After the pooling operation, the dimensions of the feature vector extracted by each convolution kernel are obviously reduced, and the most important semantic information is reserved.

E. CLASSIFICATION MODULE

The feature matrix extracted by CNN model or the semantic matrix extracted by Capsule is input into a dropout layer to prevent the over-fitting problem. During the training process, some neurons which are selected randomly in the hidden layer do not work, but they are still retained for the next input sample. The other neurons participate in the process of computation and connection. The vector matrix is input into a full connection layer for dimension reduction. Finally, the probability distribution of the sentiment category is computed by softmax activation function $y = soft\max(x)$.

The algorithm of sentiment classification by BGRU-Capsule model is described as Table 1.

IV. EXPERIMENT AND RESULTS

A. DATASETS

There are two short text datasets used in the experiment. (1) Movie Review Data,[1] published by Cornell University, is the sentiment analysis annotation corpus. (2) NLPCC 2014 Data,[2] published by Chinese Information Processing Society of China, including more than 10,000 product reviews in English. The data distribution is shown in Table 2.

The data is initialized as lexical vector by Glove and the dimension is set to 200. The hidden vector size of unidirectional and bidirectional GRU is set to 200 and 400 respectively. The convolution window size is set to 2 and 3 respectively, which can extract both the bigram and trigram features. Both the length of attention and number of capsule are set to the sentence length. The parameter of dropout is set

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data/
[2] http://tcci.ccf.org.cn/conference/2014/pages/page04_sam.html

TABLE 1. Sentiment classification algorithm by capsule-based model.

Input: Short text Data
Output: The probability distribution of the sentiment category.

1. Represent each short text in training data as a sequence of pre-trained word embedding *Seq*.
2. Truncate or pad every *Seq* with limited length of 50 words.
3. **for** epoch = 1…N **do**
   - 3.1 Select a random mini-batch containing M training samples
   - 3.2 **for** training sample i = 1…M **do**
     - ➢ Feed training sample to bidirectional BGRU and get the output $BG = \{bg_1^*, ..., bg_i^*, ..., bg_n^*\}$.
     - ➢ Connect with a Capsule layer by dynamic Routing algorithm

$$\hat{u}_{o|i} = W_{io} bg_i^*$$

Procedure Routing($\hat{u}_{o|i}$, r_num)

Begin

for all capsule i in input layer and capsule o in output layer: $b_{io} \leftarrow 0$.
for r_num iterations do
for all capsule i in input layer: $c_{io} \leftarrow \text{softmax}(b_{io})$
for all capsule o in output layer: $s_o \leftarrow \sum_i c_{io} \hat{u}_{o|i}$
for all capsule o in output layer: $v_o \leftarrow \text{squash}(s_o)$
for all capsule i in input layer and capsule o in output layer: $b_{io} \leftarrow b_{io} + \hat{u}_{o|i} \cdot v_o$

End procedure

- ➢ Calculate the probability distribution of sentiment category by *softmax*.

   end for
end for

TABLE 2. Distribution of two datasets.

	Training Set	Test Set	Total
Movie Review Data	9594	1068	10662
NLPCC2014	10000	2500	12500

to 0.5 and batch-size is 64. Adam is the optimized approach with learning rate of 0.001.

B. EVALUATION ON MOVIE REVIEW DATA

Different modules are assembled for comparison, including BGRU module, attention module, CNN module and Capsule module. Three activation functions are used on different assembling models, namely *sigmoid*, *relu* and *tanh*. The results evaluated by accuracy metric are shown in Table 3.

It can be seen from Table 3 that our BGRU + Capsule model achieves the best performance on different activation functions. It outperforms the other 4 models. Especially, *tanh* function gets the best result of 0.8255. It is mainly due to the reason that *tanh* function can alleviate the gradient disappearing problem with the increasing iteration number.

Capsule network combines the advantages of CNN and attention. CNN mainly focuses on extracting the local features of text and attention mechanism is more concerned with information in the context. The number of capsules is set to the sentence length. Each capsule can capture the features of word by the dynamic routing algorithm iteratively, and also the semantic information within the context is retained by the learning model. And so our BGRU + Capsule model with the simple structure achieves the best performance. At the same time, the parameter c of the capsule network does not participate in the reverse propagation. Compared with the complex model, it can reduce the complexity of model training and shorten training time greatly.

We also compare the performance of our best model BGRU + Capsule with other related work on Movie Review Data as following.

RAE [28]**:** A model based on recursive auto encoders for sentence-level prediction of sentiment label distribution, which learns vector space representations for multi-word phrases.

LR-Bi-LSTM [29]**:** A model uses Bi-LSTM with linguistic regularizer and it outperforms Bi-LSTM significantly.

TABLE 3. Result comparison by different assembling models on movie review data.

Model	Accuracy		
	sigmoid	*relu*	*tanh*
BGRU+Attention+CNN+Capsule	0.7899	0.7908	0.7936
BGRU+Attention+CNN	0.7917	0.8040	0.8200
BGRU+Attention+Capsule	0.7927	0.8039	0.8124
BGRU+CNN	0.7880	0.7861	0.7900
BGRU+Capsule	**0.8011**	**0.8152**	**0.8255**

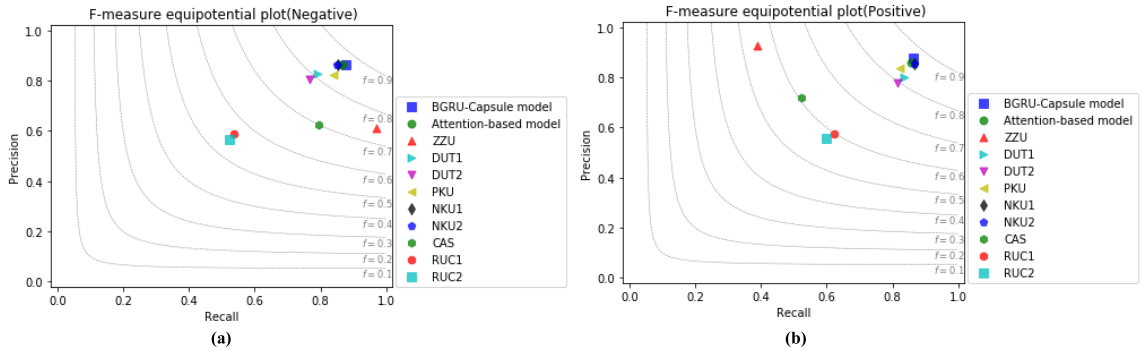


FIGURE 3. F-measure equipotential plot of different solutions in NLPCC2014 Task II. (a) F-measure equipotential plot (Negative). (b) F-measure equipotential plot (Positive).

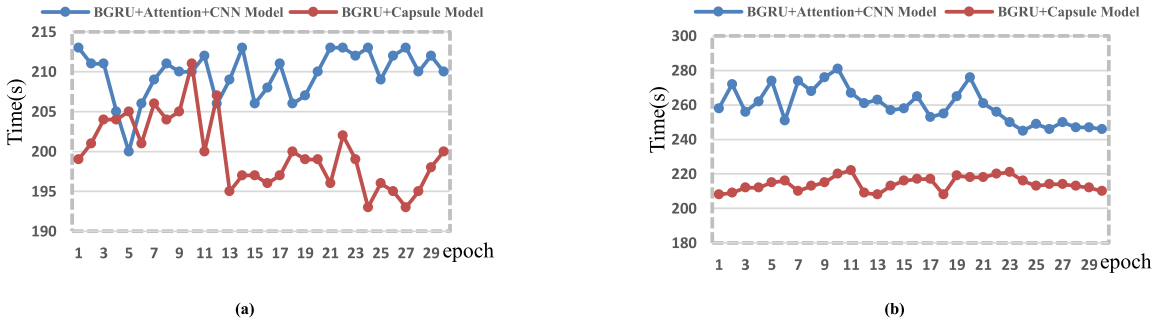


FIGURE 4. Time Performance Comparison on Two DataSets. (a) Movie Review Data. (b) NLPCC Data.

TABLE 4. Result comparisons with other methods on movie review data.

Model	Accuracy
RAE [28]	0.7770
LR-Bi-LSTM [29]	0.8210
CNN-multichannel [5]	0.8110
CNN-non-static [5]	0.8150
TE-LSTM$_{+c,p}$ [30]	0.8220
Capsule-A [19]	0.8130
Capsule-B [19]	0.8230
BGRU-Capsule	**0.8255**

CNN-multichannel [5]**:** A model with two sets of word vectors. Each set of vector is treated as a 'channel' and each filter is applied to both two channels. The model is able to fine-tune one set of vectors while keeping the other static.

CNN-non-static [5]**:** A model with pre-trained vectors by word2vec. The pre-trained vectors are fine-tuned for each task.

TE-LSTM$_{+c,p}$ [30]**:** A model proposed to utilize POS tags to control the gates of tree-structured LSTM networks, which is combined with representations of phrases.

Capsule-A [19]**:** A model adopts only one parallel network with filter windows size of 3 in the convolutional layer.

Capsule-B [19]**:** A model adopts three parallel networks with filter windows size of 3, 4, 5 in the N-gram convolutional layer.

The performance comparisons are shown in Table 4.

It can be seen from Table 4 that our BGRU-Capsule Model performs best, which fuses the advantage of BGRU and Capsule network to accomplish the task.

C. EVALUATION ON NLPCC DATA

The BGRU-Capsule Model is also implemented on the NLPCC 2014 data to further verify the model's effectiveness. Here, the convolutional window size in CNN is set to 2 and 3 respectively and the corresponding model is labeled as CNN_2 and CNN_3. The experimental results on 5 different models are shown in Table 5.

The results in Table 5 show that our BGRU + Capsule model outperforms the other models on *relu* and *tanh* activation functions. In addition, it is found that these different models perform better on Negative data than Positive data mostly. The evaluation metrics of Precision, Recall and F-measure are used to compare our model with NLPCC2014 Task II result,[3] which are shown in Table 6 and Fig.3.

As can be seen from Table 6 and Fig.3, our model performs better than most of the participating systems of NLPCC2014 Task II, and especially achieves a better performance than the best result on Negative data in precision, recall and F-measure.

D. ANALYSIS OF TIME PERFORMANCE

The structure of BGRU + Capsule model is relatively simple. The internal parameters of the network do not participate in the reverse propagation, which reduces the training time greatly. In order to further verify the efficiency of the Capsule network, we conduct time comparison experiments by

[3]http://tcci.ccf.org.cn/conference/2014/index.html

TABLE 5. sentiment classification results by different models on NLPCC data (accuracy).

	Sigmoid		*Relu*		*Tanh*	
	Pos	Neg	Pos	Neg	Pos	Neg
BGRU	0.7000	0.7680	0.7704	0.7824	0.7568	0.8120
$CNN_2 + CNN_3$	0.7824	0.8040	0.7608	0.7824	0.7640	0.7832
BGRU + CNN_2 + CNN_3	0.7904	0.8176	0.7688	0.7576	0.7904	0.8224
BGRU+Attention+CNN	**0.8560**	**0.8672**	0.8344	0.8368	0.8064	0.8664
BGRU+Capsule	0.8544	0.8648	**0.8568**	**0.8667**	**0.8640**	**0.8784**

TABLE 6. Comparison result with NLPCC2014 Task II.

	Positive			**Negative**		
	Precision	Recall	F-measure	Precision	Recall	F-measure
BGRU+Capsule model	**0.8754**	0.8640	**0.8697**	**0.8659**	**0.8784**	**0.8721**
NLPCC2014 Task II Medium(F_1)	0.7790	0.8160	0.7970	0.8070	0.7690	0.7870
NLPCC2014 Task II Best（F_1）	0.8560	**0.8660**	0.8610	0.8640	0.8550	0.8600

BGRU + Attention + CNN model and the BGRU + Capsule model separately. The number of iterations is set to 30 and the results are shown in Fig.4. As can be seen, the BGRU + Capsule model has less training time mostly by different epoch. Especially, it performs more stable on NLPCC dataset.

V. CONCLUSIONS

We propose a novel capsule-based hybrid model for the short text sentiment classification task, which not only makes full use of the forward and backward information of short texts and shortens the distance between the interdependent features, but also can extract richer textual information to improve the expression ability, including word position, semantic and syntactic structure. Especially, the capsule based model has the advantage of less training time and simple network structure to achieve a better performance. Simultaneously, we also devise an attention-based model for comparison. The experimental results show that the capsule-based hybrid model outperforms other related approaches on Movie Review Data and also gets a better result than the participated systems in NLPCC2014 task II. During the process of dynamic iteration, the model can adjust the attention weights smoothing and mine more hidden characteristic information. Hence, the capsule-based model can achieve high-quality features by combining the advantages of BGRU, attention mechanism and CNN.

In the future, we will apply our model to other related sentiment analysis tasks, such as aspect-based opinion classification and cross-domain sentiment analysis. Furthermore, more auxiliary neural networks will be explored to enforce our model for higher performance.

REFERENCES

[1] B. Liu, "Sentiment analysis and opinion mining," *Synth. Lectures Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, May 2012.

[2] Z. Y. Yan, Q. Bin, S. Q. Hui, and L. Ting, "Large-scale sentiment lexicon collection and its application in sentiment classification," *J. Chin. Inf. Process.*, vol. 31, no. 2, pp. 187-193, 2017.

[3] Odbal and Z. F. Wang, "Emotion analysis model using Compositional Semantics," *Acta Automatica Sinica*, vol. 41, no. 12, pp. 2125–2137, 2015.

[4] P. Li, W. Xu, C. Ma, J. Sun, and Y. Yan, "IOA: Improving SVM based sentiment classification through post processing," in *Proc. 9th Int. Workshop Semantic Eval.*, Jun. 2015, pp. 545–550.

[5] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2014, pp. 1746–1751.

[6] Y.-X. He, S.-T. Sun, F.-F. Niu, and F. Li, "A deep learning model enhanced with emotion semantics for microblog sentiment analysis," *Chin. J. Comput.*, vol. 40, no. 4, pp. 773–790, Apr. 2017.

[7] C. N. D. Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proc. COLING 25th Int. Conf. Comput. Linguistics, Tech. Papers*, Aug. 2014, pp. 69–78.

[8] J. Deriu, M. Gonzenbach, F. Uzdilli, A. Lucchi, V. De Luca, and M. Jaggi, "SwissCheese at SemEval-2016 Task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision," in *Proc. 10th Int. Workshop Semantic Eval.*, Jun. 2016, pp. 1124–1128.

[9] J. Xu, D. Chen, X. Qiu, and X. Huang, "Cached long short-term memory neural networks for document-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Nov. 2016, pp. 1660–1669.

[10] G. Zhou, J. Zhao, and D. Zeng, "Sentiment classification with graph co-regularization," in *Proc. COLING 25th Int. Conf. Comput. Linguistics, Tech. Papers*, Aug. 2014, pp. 1331–1340.

[11] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," in *Proc. COLING 26th Int.l Conf. Comput. Linguistics, Tech. Papers*, Dec. 2015, pp. 3298–3307.

[12] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.*, Jul. 2015, pp. 1556–1566.

[13] P. Zhou *et al.*, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Aug. 2016, pp. 207–212.

[14] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "ABCNN: Attention-based convolutional neural network for modeling sentence pairs," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 259–272, Dec. 2016.

[15] D. Bahdanau, K. Cho, and Y. Bengio. (Sep. 2014). "Neural machine translation by jointly learning to align and translate." [Online]. Available: https://arxiv.org/abs/1409.0473

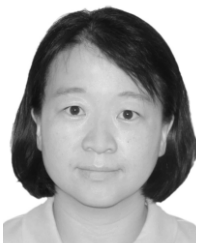
[16] T. Yanase, K. Yanai, M. Sato, T. Miyoshi, and Y. Niwa, "Bunji at SemEval-2016 task 5: Neural and syntactic models of entity-attribute relationship for aspect-based sentiment analysis," in *Proc. SemEval*, Jun. 2016, pp. 289–295.

[17] A. Galassi, M. Lippi, and P. Torroni. (Feb. 2019). "Attention, please! A critical review of neural attention models in natural language processing." [Online]. Available: https://arxiv.org/abs/1902.02181

[18] S. Sabour, N. Frosst, and G. E. Hinton. (2017). "Dynamic routing between capsules." [Online]. Available: https://arxiv.org/abs/1710.09829

[19] W. Zhao, J. Ye, M. Yang, Z. Lei, S. Zhang, and Z. Zhao. (2018). "Investigating capsule networks with dynamic routing for text classification." [Online]. Available: https://arxiv.org/abs/1804.00538

[20] J. Kim, S. Jang, S. Choi, and E. Park. (2018). "Text classification using capsules." [Online]. Available: https://arxiv.org/abs/1808.03976
[21] N. Zhang, S. Deng, Z. Sun, X. Chen, W. Zhang, and H. Chen, "Attention-based capsule networks with dynamic routing for relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Nov. 2018, pp. 986–992.
[22] B. Zhang, X. Xu, M. Yang, X. Chen, and Y. Ye, "Cross-domain sentiment classification by capsule network with semantic rules," *IEEE Access*, vol. 6, pp. 58284–58294, Oct. 2018.
[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[24] X. Zhou, X. Wan, and J. Xiao, "Attention-based LSTM network for cross-lingual sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Nov. 2016, pp. 247–256.
[25] M. Sundermeyer, T. Alkhouli, J. Wuebker, and H. Ney, "Translation modeling with bidirectional recurrent neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2014, pp. 14–25.
[26] A. Vaswani *et al.* (2017). "Attention is all you need." [Online]. Available: https://arxiv.org/abs/1706.03762
[27] G. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with EM routing," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, Feb. 2018, pp. 1–29.
[28] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-supervised recursive autoencoders for predicting sentiment distributions," in *Proc. Conf. Empirical Methods Natural Lang. Process. Assoc. Comput. Linguistics*, Jul. 2011, pp. 151–161.
[29] Q. Qian, M. Huang, J. Lei, and X. Zhu. (Nov. 2016). "Linguistically regularized LSTMs for sentiment classification." [Online]. Available: https://arxiv.org/abs/1611.03949
[30] M. Huang, Q. Qian, and X. Zhu, "Encoding syntactic knowledge in neural networks for sentiment classification," *ACM Trans. Inf. Syst.*, vol. 35, no. 3, Jun. 2017, Art. no. 26.



YONGPING DU received the Ph.D. degree in computer science from Fudan University, Shanghai, China, in 2005. She is currently an Associate Professor with the Beijing University of Technology. Her research interests include information retrieval, information extraction, and natural language processing.

XIAOZHENG ZHAO received the B.S. degree in computer science and technology from the Beijing University of Technology, China, in 2016, where she is currently pursuing the M.S. degree in computer science and technology. Her research interests include natural language processing and sentiment analysis.

MENG HE received the B.S. degree in computer science and technology from the Qingdao University of Science and Technology, in 2017. She is currently pursuing the M.S. degree in computer science and technology with the Beijing University of Technology, China. Her research interests include natural language processing and cross-domain sentiment analysis.

WENYANG GUO received the B.S. degree in computer science and technology from the Beijing University of Technology, China, in 2018, where she is currently pursuing the M.S. degree in computer science and technology. Her research interests include natural language processing and question answering.