PROCEEDINGS OF SCIENCE

A Hybrid Neural Network for Sentence Classification

Xiaoping Du[1]
School of Software, Beihang University
Beijing, 100191, China
E-mail: xpdu@buaa.edu.cn

Renquan Zhou[2,3]
School of Software, Beihang University
Beijing, 100191, China
E-mail: zhourenq123@163.com

The sentence classification is the foundation of many Natural Language Processing applications. Prior neural network which use one type network for sentence classification can't use the abundant information in a sentence. In this paper, we proposed a hybrid neural network in combination with recurrent neural network and convolutional neural networks for sentence classification. The recurrent neural network can model long distance global information in a text, but it can't effectively extract the local information and convolutional neural network inversely. The proposed hybrid neural network takes full advantage of the advantages of these two networks while extracting global feature and local feature at the same time. In order to get the global feature, we also proposed three different methods to make use of hidden states generated by recurrent neural network. We conducted experiments on four public open datasets. The results show that our hybrid neural network does better than models by using the recurrent neural network or the convolutional neural networks alone, higher and completive classification accuracy is obtained.

CENet2017
22-23 July, 2017
Shanghai, China

[1]Supported by the National Natural Science Foundation of China (No.71673276)
[2]Speaker
[3]Corresponding Author

PoS(CENet2017)057

1.Introduction

With the development of deep learning, deep neural network has demonstrated great capability in various tasks of Natural Language Processing, such as text classification, machine translation and other fields [1]. The sentence classification is the foundation of many Natural Language Processing applications, such as question classification and sentiment analysis [2]. According to the number of words that each record contains, the text can be divided into sentence, paragraph and document. This study is based on sentence, which contains dozens of words.

Kim [3] first applied the convolution neural network for text classification by using convolution filter to extract the local feature with different widths, then used the pooling layer (Max-pooling) to get text vector representation and proved that the convolution neural network can get considerable or better results than traditional feature engineering method. Kalchbrenner[4] proposed a convolutional neural network for text classification too while the pooling method was changed to dynamic K-Max poolingto extract multiple ordered features simultaneously and use some global information of the whole text.

The recurrent neural network is widely used in the sequence-sequence language tasks. Bahdanau[5] applied the recurrent neural network to Machine Translation tasks. Bi-LSTM model was used to encode source language and decode target language to get better results than the feature-based Machine Translation methods. Zichao Yang[6] proposed the hierarchical attention network for document classification. The first layer used Bi-LSTM and attention mechanism to get sentence vector representation with processing word vectors one by one, the second layer used the same structure as the first layer to obtain document vector representation sentence by sentence too. Attention mechanism can focus on core words and sentences in a document.

There are some studies about text classification by using the model with conjunction of recurrent neural network and convolutional neural network. Chunting Zhou [7] proposed a network called C-LSTM for text classification which used convolutional neural network for feature extraction firstly, and then used the LSTM network to generate text representation, the results were better than that by using convolutional neural network alone. Siwei Lai[8] used recurrent convolution neural network for text classification. The recurrent network was used to capture contextual information of words in text from left and right respectively as far as possible, and used the convolutional network for feature extraction to generate document representation. This model not only used the information above a word but used the information below.

At present, most proposed model have used only one network for text classification, either recurrent neural network or convolutional neural network. Even if the combined model uses one network as the main feature extractor, the other is just an auxiliary. Intuitively, the recurrent neural network is more conducive to model the global information. Whereas the convolutional neural network has advantage for modeling local information, we can learn better representation for text with fully utilization of local information and global information together to improve the text classification results. This paper proposed a hybrid neural network by using both two type networks for sentence classfication. This paper contributes mainly from three aspects: 1) it proposes a hybird neural network in combination with CNN and Bi-LSTM for sentence calssfication; 2) it proposes three effective methods to get global feature from the hidden states

generated by Bi-LSTM; 3) it designs and conducts experiments on public open dataset to study the performance of proposed hybrid model.

2.The Model

In this paper, a hybrid neural network based on convolutional neural network (CNN) and recurrent neural network (we use Bi-LSTM in this study) is proposed for text classification. The structure of the proposed model is shown in Figure 1. The input of the model is the word sequence of the text, which is pre-processed as token-ids (padding the text to the Max-length if the length is short); the output of the model contains C values representing the probability of text belongs to each category. The category with the maximum probability is the final results of the text classification. Each token-id in the sentence is expressed as real-value vector, and then the CNN and Bi-LSTM is used for feature extraction, CNN for local feature and Bi-LSTM for global feature respectively. The output of Bi-LSTM is a sequence of hidden states. THus we proposed three different methods to extract global feature from the hidden states. Finally, the local feature and global feature are connected as the representation of sentence, followed by a Drop-out layer and a full-connect layer. The output is used as the input of the Softmax classier.

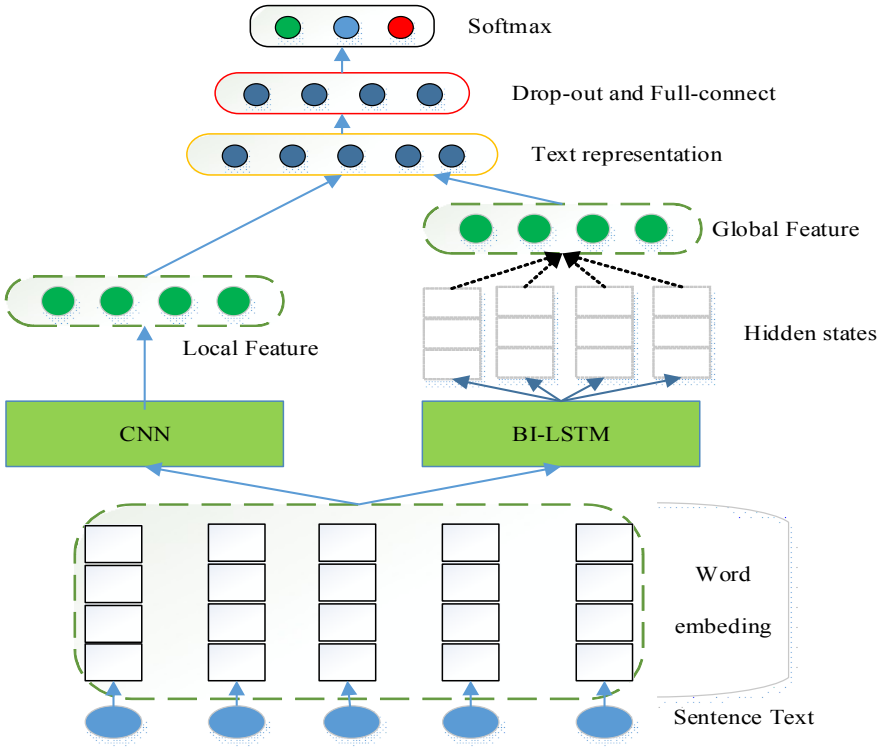

Figure 1: Structure of the Proposed Hybrid Neural Network

2.1Text Representation

Each sentence S contains a sequence of words $X_{1:n}$ and each word in sentence is represented as a d-dimensional real-value vector, define $S= X_{1:n} \in R^{n*d}$. Pre-trained word2vec[9] can improve the classification performance because the pre-trained wrod2vec uses extra unlabeled corpus, belonging to the concept of transfer learning strictly speaking. The core of our study focuses on the representation ability of hybrid model, so the words in all models do not use pre-trained word vector but using the random real-value for initialization.

In order to facilitate the processing of model, we pad the sentence to the maximum length with zero if the length is less than the maximum sentence length .

2.2Local Feature

The CNN is used to extract local feature. CNN extracts features by using convolution filter $W \in R^{h*d}$ refer to Kim. The convolution filter i is used to continuously words of a sentence with width h. For example, a feature window $X_{j:j+h-1}$ produces convolution features C_{ij} as follow,

$$C_{ij} = f\left(W * X_{j:j+h-1} + b\right) \tag{2.1}$$

Here $b \in R$ is the bias, f is a nonlinear function, Relu is chose in this paper, which shows good results in many networks. Finally, the features extracted by each convolution filter can be represented as,

$$C_i = [C_{i1}, C_{i2}, \ldots, C_{i(n-h+1)}] \tag{2.2}$$

Every convolution filter is applied to all possible windows to produce n-h+1 features. In order to extract features from multiple views, a number of different filters H (with different h) are used, each type filter uses n times respectively.

A convolution filter may extract a lot of useless features because the semantic of sentence is usually concentrated to a few words and phrases. The Max-pooling is used behind the convolutional layer to extract the most useful feature C_{imax} of each convolution filter, and finally get the local feature of sentence denoted as S_{local}, $k=|H|*n$ is the total number of all convolution filters.

$$S_{local} = [C_{1max}, C_{2max}, \ldots, C_{kmax}] \tag{2.3}$$

2.3Global Feature

The recurrent neural network is used to extract global information of a sentence. The recurrent neural network have problems in semantic bias and hard to train. LSTM uses the hidden state to preserve the intermediate information, making it easy for model training[10]. Based on the representation learning from two directions (called Bi-LSTM). it can effectively solve the problem of semantic bias. The update process of hidden state hi of LSTM in each direction is,

$$h_i = F(X_i, h_{i-1}) \tag{2.4}$$

Where X_i is the current word vector in processing, h_{i-1} represents the previous state and h_i the current state. The state is denoted as a real-value vector $h_i \in R^k$,in which, k refers to the dimension of hidden state. F stands for a series of LSTM operations, abbreviated as F. The output of Bi-LSTM is a sequence of hidden states as follows,

$$S_{hidden} = [hl_{1:n}, hr_{1:n}] \tag{2.5}$$

Where $hl_{1:n}$ represents the sequence of n states generated by the forward-LSTM and $hr_{1:n}$ in the inverse direction.

After obtaining the state sequence, we propose three methods to generate the final global feature. The structure is shown in Figure 2, referring to three variants of hybrid neural network.

Method one: It simply uses the last state of the sequence as the global feature of the sentence, and hl_{last} and hr_{last} is the last state of the forward and backword state sequence.the model can be abbreviated as Hybrid-1.

$$S_{global} = [hl_{last}, hr_{last}] \tag{2.6}$$

Method two: Using the average value of all hidden states in sequence as the global feature the sentence, we define hl_{avg} and hr_{avg} as the element-wise mean of all states, the globasl feature is the connection of hl_{avg} and hr_{avg} same as formula 2.6. The model is called Hybrid-2.

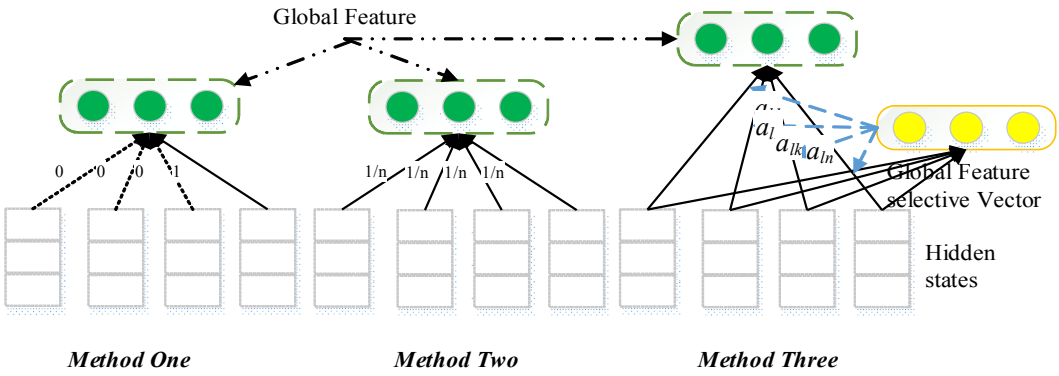

Figure 2: Three Proposed Methods to Generate Global Geature

Method three: we adopt global attention mechanism to "concern" the most important states in a sequence and assign different states with different weights. The model is called Hybrid-3. For the forward state sequence, the forward-direction global feature hl_{atten} of the sentence is computed as follows,

$$al_i = \frac{exp(hl_i * u_g)}{\sum_{j=1}^{n} exp(hl_j * u_g)} \tag{2.7}$$

$$hl_{atten} = \sum_i al_i hl_i \tag{2.8}$$

Among the formulas above, al_i represents the weighted weight of the state hl_i in the global feature. Similarly, the inverse-direction hr_{atten} can be computed in the same way.

u_g represents a global feature selection vector, which is the output of a multilayer perceptron with local feature generated by CNN as the input. Use the local feature to generate the feature selection vector because the local feature represents which feature is important for a sentence; therefore, it can be used as feature selection metric. The u_g is calculated as follows,

$$u_g = W_g * S_{local} + b_g \tag{2.9}$$

we don't use local feature as the selection vector directly because the attention mechanism requires the selection vector must has the same dimension k with the hidden state of Bi-LSTM. It is difficult to train and adjust the model while using it directly, which makes the model highly coupled. u_g can be seen as a representation of local feature in another feature space. After get hl_{atten} and hr_{atten}, the global feature S_{global} is the connection of them same as formula 2.6.

2.4Text Classification

Given the global and local feature of the sentence text, the text representation S_{txt} is the connection of them. A linear transformation layer and Softmax layer are added on the top of the model to produce probabilities over the class space C. To avoid overtting, dropout with a probability q is applied to the penultimate layer during training (disabled when testing).

Where $\odot$ is an element-wise multiplication operator, W_q is the masking vector with drop rate q, and W_c , b_c are variables. P_c is the probability of the sentence classified to class c, and the class c with the maxmium probability is the classification result of a sentence. As our model is a supervised method and each sentence has a golden category, the categorical cross-entropy is used as the objective function during training.

$$S_{txt} = [S_{local}, S_{global}] \tag{2.10}$$

$$y = W_c * (S_{txt} \odot W_q) + b_c \tag{2.11}$$

$$P_c = \frac{exp\quad(y_c)}{\sum_{c' \in C} exp\quad(y_{c'})} \tag{2.12}$$

3.Experiment Setting

3.1Datasets

We conduct experiments to verify the performance of the proposed Hybrid model based on public open datasets. The statistics of the used datasets are shown in Table 1 as following.

Datasets	C	Max-length	Avg-length	N	|V|	Test
MR	2	59	20	10662	18765	CV
TREC	6	33	10	5952	9592	500
SST-1	5	52	18	9613	16185	2210
SST-2	2	52	19	9613	16185	1821

Table 1: Summary statistics for the datasets. C: Number of target classes. Max-length: Max sentence length. Avg-length: Average sentence length. N: Dataset size. |V|: Vocabulary size. Test: Test set size (CV means there was no standard train/test split and thus 10 fold CV was used).

MR: Movie review dataset with one sentence for each review. The classification tasks is to classify comments into positive/negative reviews.

TREC: TREC problem sets, the task is to classify a problem into six problem categories (whether the question is about people, location and numbers, etc.).

SST-1: Stanford Sentiment Teeebank--An extension of the MR dataset but has a splits of train/verification/test and fine-grained labels (very positive, positive, neutral, negative, very negative), re-labeled by Socher et al[11].

SST-2: The dataset is the same as SST-1, but it doesn't contain neutral data and has only two categories: positive/negative.

3.2Baselines and Parameters

We mainly study the ability of the proposed hybrid neural network for modeling the sentence text. The basic neural network is used as baselines and contrast model.

CNN-standard (Model Ⅰ), a standard CNN network proposed by Kim. Bi-LSTM-Last (model Ⅱ) model, which uses Bi-LSTM to get the representation of sentence and use the last state as the feature. Bi-LSTM-Avg (model Ⅲ), same as Bi-LSTM-Last model but use the average of all hidden states as the feature.

We program and implement our proposed model on the platform Tensorflow0.12, a famous deep learning framework with great flexibility and capacity. The parameter setting in convolution neural network for all above models refer to Kim's paper, the convolution filter widths H are set to [2, 3, 4]. Each width has a set of 100 convolution filters.We conduct grid search on MR datasets and find that Model III can get a good performance when the hidden state dimension k in Bi-LSTM is set to 100, the dimension of Bi-LSTM and the global feature selection vector is set to 100 during training.

We do not otherwise perform any dataset-specific tuning other than early stopping on dev sets. Drop the rate default set to 0.5 and the training is done through AdamGrad learning method with learning rate 1e-3. In addition, a l_2 norm constraint of the weights W_c (and W_g for Hybrid-3 model)is imposed during training as well.

4.Results and Discussion

The accuracy is used as the evaluation metric, the classification accuracy of all above model onall dataset is shown in Table 2.

model	MR	TREC	SST-1	SST-2
I	76.61	91.2	41.34	80.45
II	76.59	88.4	41.12	80.31
III	74.62	82.6	37.13	76.14
Hybrid-1	77.51	**92.3**	**41.77**	80.8
Hybrid-2	**77.6**	**92.3**	41.43	**81.1**
Hybrid-3	77.34	89.4	41.33	80.39

Table2: Classification Accuracy of All Models

dataset	Model	I	II	III
MR	**Hybrid-1**	+1.09	+1.20	+3.87
	Hybrid-2	+1.29	+1.32	+3.99
	Hybrid-3	+0.95	+0.98	3.64
TREC	**Hybrid-1**	+1.20	+4.41	+11.74
	Hybrid-2	+1.20	+4.41	+11.74
	Hybrid-3	-1.97	+1.13	+8.23
SST-1	**Hybrid-1**	+1.04	+1.58	+14.62
	Hybrid-2	+0.22	+0.75	+11.58
	Hybrid-3	-0.02	+0.51	+11.31
SST-2	**Hybrid-1**	+0.44	+0.61	+5.70
	Hybrid-2	+0.81	+0.98	+6.51
	Hybrid-3	-0.07	+0.10	+5.58

Tabel 3: Relative accuracy improvements

According the results, we can find that the proposed Hybrid neural network do better than the model with the single convolutional neural network CNN or recurrent neural network Bi-LSTM. The relative accuracy improvements are listed in Table 3. Take the MR dataset for example and make comparison with CNN-standard, Hybrid-1, Hybrid-2 and Hybrid-3 to get accuracy improvement of 1.09%, 1.29%, 0.95% respectively, 1.20%, 1.32%, 0.98% compared to Bi-LSTM-Last and 3.87%, 3.99%, 3.64% to Bi-LSTM-Avg. The reason for the results is that the convolution neural network is good at local information extraction and recurrent neural network at global information extraction. The hybrid neural network can effectively utilize global information and local information of the sentence at the same time to get a better representation of the text, which is consistent with our hypothesis.

The convolutional neural network is superior to the recurrent neural network for shorter sentence. From the results, Model I and Model II get similar result in dataset MR and SST, but Model I is significantly superior to Modle II in TREC (91.2% vs 88.4%). We analyse it because the sentence in TREC is shorter, the *Max-length* and *Avg-length* is 33 and 10, which is less than three other datasets. The semantic information of a sentence in TREC is concentrate upon a few of words or phrases in accordance with the characteristics of convolutional neural network, which makes Model I better. With the growth of the sentence length, the recurrent neural network can become better gradually, which is more suitable for global extraction.

Based on the results, we can also find that three hybrid models proposed in this paper show little difference (77.51%, 77.6%, 77.34 on MR dataset respectively). The Hybrid-1 model is slightly better than the Hybrid-2 and Hybrid-3 models. Due to the characteristic of short sentence text, we guess that this phenomenon, which makes the local feature extracted by convolutional neural network plays a major role in the classification results. With the increase of the length of sentence, the performance enhancement speed of Hybrid-2 and Hybrid-3 model is faster than Hybird-1, which can be seen from the performance in MR dataset contrast to TREC.

In order to estimate the influence of the dimension k setted in Bi-LSTM, we conduct experiments on MR dataset with the model Hybrid-2. The result is shown in Figure 3. The result

shows that classification results will be worse if the k is too small or too large, while the impact is not particularly significant.

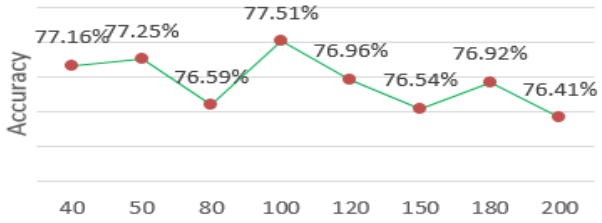

Figure 3: Influence of Bi-LSTM Dimension k

5.Conclusion

This paper proposed a hybrid neural network for sentence classfication and three different methods to extract global feature from the hidden states generated by Bi-LSTM. The network can use the global and local information of a sentence at the same time and get improvements on classfication accuray compared to prior single model.

The model also has a lot of space for improvement. For example, the current work is done on the sentence text and it needs to further study the paragraph, the document text and study the influence of sentence length; three variants of hybrid model have been proposed in this paper. More and better models need further exploration.

References

[1] Lopez M M, Kalita J. *Deep Learning applied to NLP*[J]. 2017.

[2] Zhao Z, Wu Y. *Attention-based Convolutional Neural Networks for Sentence Classification*[C]// INTERSPEECH 2016, Conference of the International Speech Communication Association. 2016.

[3] Kim Y. *Convolutional Neural Networks for Sentence Classification*[J]. Eprint Arxiv, 2014.

[4] Kalchbrenner N, Grefenstette E, Blunsom P. *A Convolutional Neural Network for Modelling Sentences*[J]. Eprint Arxiv, 2014, 1.

[5] Bahdanau D, Cho K, Bengio Y. N*eural Machine Translation by Jointly Learning to Align and Translate*[J]. Computer Science, 2014.

[6] Yang Z, Yang D, Dyer C, et al. *Hierarchical Attention Networks for Document Classification*[C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016:1480-1489.

[7] Zhou C, Sun C, Liu Z, et al. *A C-LSTM Neural Network for Text Classification*[J]. Computer Science, 2015, 1(4):39-44.

[8] Lai S, Xu L, Liu K, and Zhao J. *Recurrent convolutional neural networks for text classification*. In Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[9] Mikolov T, Sutskever I, Chen K, et al. *Distributed Representations of Words and Phrases and their Compositionality*[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.

[10] Gers F A, Schmidhuber J, Cummins F. *Learning to forget: continual prediction with LSTM*[J]. Neural Computation, 2000, 12(10):2451.

[11] Socher R, Perelygin A, Wu J Y, et al. *Recursive deep models for semantic compositionality over a sentiment treebank*[J]. 2013.