Full Attention-Based Bi-GRU Neural Network for News Text Classification

Introduction

Text classification is an important task in natural language processing (NLP), whose purpose is to assign predefined category or categories to the given text. The applications of text classification include: spam detection, sentiment analysis, topic classification and so on. With the development of the Internet and information technology, data resource becomes more massive. In order to meet the needs of a large number of news users, it's urgent to effectively manage and utilize the news. This paper aims at assigning a predefined category to the given news text. Traditional approaches of text classification first represent text with sparse lexical features, and then use a linear model or kernel method to assign label or labels to the given text. However, these methods suffer from the problem of data sparseness. In recent years, several works learn the representation of text through neural networks. For example, convolutional neural network (CNN) is used to extract features from information of fixed-size windows, and recurrent neural network (RNN) based on long short-term memory (LSTM) is employed to obtain further context information. These methods have better performance than traditional approaches. Moreover, in order to enable the model to utilize the key information, Yang introduced attention mechanism and proposed hierarchical attention networks (HAN), which improved the effect of text classification. However, due to that HAN only assigns weights to the whole outputs of the encoder once, it fails to get richer key information for each step. In this paper, we focus on news text classification. In order to pay more attention to key information, a bidirectional GRU (Bi-GRU) neural network based on full attention mechanism is proposed, which is called FABG. The Bi-GRU is used to extract context information, and full attention mechanism is used to focus on key information. Different from traditional attention mechanism that calculates the representation of the whole text, FABG re-calculates the vectors of each step by assigning weights to the encoded outputs of current and previous steps, which is full attention mechanism. Experiments show that FABG achieves good results in the text classification of news topics. The main contributions in this work are: a model with full attention mechanism is proposed, which can obtain richer key information. And experiments on two news datasets show that this model outperforms the baselines, which proves the effectiveness of this model. Also, a Chinese news dataset is built with higher timeliness by collecting news in www.chinanews.com from January to February 2019.

Full attention-based bi-gru neural network

In the full attention-based Bi-GRU neural network, the Bi-GRU is used to learn the latent information. And full attention mechanism is used to strengthen the influence of the key information on the representation of each step. Pooling layer extracts the features required for classification. The architecture is shown in the following figure:

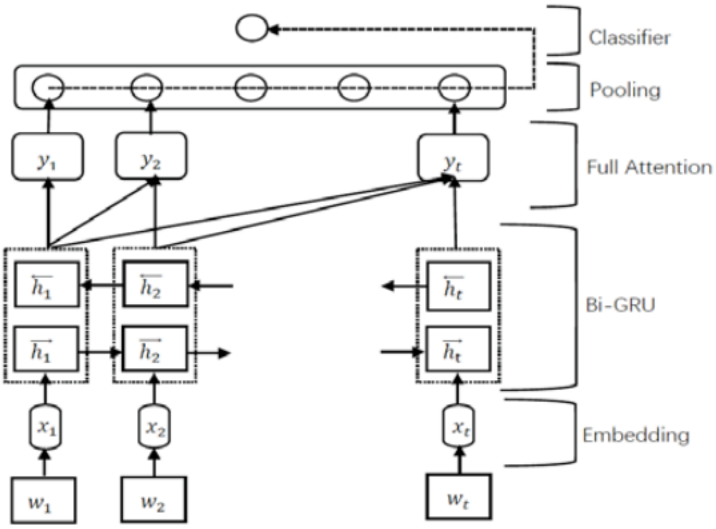

Figure 1. The architecture of FABG.

First, by using the word embedding layer, the word w_i is converted into the vector x_i whose size is fixed. And then the new representation h_i of the i-th step is learned by the Bi-GRU. Next, the outputs of the Bi-GRU are put into the full attention layer to obtain a representation with richer key information. The final text representation is extracted by the pooling layer. At last, the probability distribution of the categories is calculated by the classifier.

A. **Word Embedding**

Word is the basic unit in FABG. In word embedding layer, we use a R^{N*d} dictionary, where N is the number of words in the dictionary and d is the dimension of the vectors. Given a piece of text consisting of T words, the t-th word in the text is represented by w_t, and then each word is converted into a d-dimensional vector x_t , also $x_i \in R^d$. The matrix of the input text is expressed as in

$$X = [x_1; x_2; ...; x_T] \in R^{T\times d} \qquad (1)$$

B. **Encoder Based on Bi-GRU**

Encoding the semantic information by Bi-GRU, FABG can learn the semantic dependence among words. The GRU is a variant of the LSTM. It uses gating mechanism to track the state of the sequence. There are two types of gates in the GRU: the update gate and the reset gate. The update gate is used to decide how much past information is brought into current state and how much new information is added. The reset gate controls how much information of previous steps is written into current candidate state h_t . h_t is the output of the GRU at step t, h_{t-1} is the state of step t-1, and z_t represents the update gate. At step t, the calculation of the new state h_t is

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot h_t \qquad (2)$$

This is a linear interpolation between the previous state t 1 h and the current candidate state t h . Through the update gate z_t , the state of step t can obtain the status information of step t-1 and current candidate information h_t . x_t is the input of step t , the update function of z_t is

$$z_t = \sigma(W_{z_t} x_t + U_{z_t} h_{t-1} + b_{z_t}) \qquad (3)$$

r_t is the reset gate, h_t can be calculated by

$$h_t = \tanh(W_{h_t} x_t + r_t \odot (U_{h_t} h_{t-1}) + b_{h_t}) \qquad (4)$$

By the reset gate r_t , candidate state of step t can obtain the information of input t x and the status information h_{t-1} of step t-1. The update function of r_t is

$$r_t = \sigma(W_{r_t} x_t + U_{r_t} h_{t-1} + b_{r_t}) \tag{5}$$

FABG uses a Bi-GRU to get annotations of words by summarizing information from both directions for words, and therefore incorporates the contextual information in the annotation. The Bi-GRU contains the forward GRU h_t which reads the sentence from step 0 to t and the backward GRU h_t .

$$\overrightarrow{h_t} = \overrightarrow{GRU}(x_t), t \in [1,T] \tag{6}$$

$$\overleftarrow{h_t} = \overleftarrow{GRU}(x_t), t \in [T,1] \tag{7}$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{8}$$

C. **Full Attention**

In sentence-level text classification, the number of words in text is small. It's difficult to obtain more semantic information. In previous works, attention mechanism assigns high weights to the outputs of the Bi-GRU corresponding to the steps where the key words are located. But the information has been filtered through the update gate and the reset gate in the GRU, which means the key information could be forgotten. Besides, the same information has different effects on different steps. In order to obtain richer key information and strengthen the influence of the key information, FABG uses attention mechanism to learn a new representation of each step by re-calculating the vector of step t through weighted summing, which is full attention mechanism. For step t, FABG contacts the outputs of the Bi-GRU to get the input of full attention layer

$$H_t = [h_1, h_2, \ldots, h_t] \tag{9}$$

By linear layer and tanh activation function, FABG gets u_{ti} , which represents the hidden representation h_i at step t.

$$u_{ti} = \tanh(W_{ti} h_i + b_{ti}) \tag{10}$$

And then, we measure the importance of the output of the Bi-GRU as the similarity of u_{ti} with a word level context vector u_w and get a normalized importance weight a_{ti} by a softmax function. The context vector u_w is different at each step, and it can be regarded as a representation of "what is the informative word at step t". It's randomly initialized and jointly learned during the process of training

$$a_{ti} = \frac{\exp(u_{ti}^T u_{wt})}{\sum_{i=1}^{t} \exp(u_{ti}^T u_{wt})} \tag{11}$$

After that, we compute the vector t y by a weighted sum of the outputs of the Bi-GRU annotations

$$y_t = \sum_{i=1}^{t} a_{ti} h_i \tag{12}$$

D. **Pooling**

In pooling layer, we perform two different calculations, max-pooling and ave-pooling, respectively on the outputs of the full attention layer to obtain the final text representation.

E. **Classifier**

We get the probability distribution by

$$p=softmax(W_cY+b_c) \qquad (13)$$

Finally, we utilize the cross-entropy loss function to calculate the loss between the real distribution q and the predicted distribution p

$$\Lambda = -\sum_i q(i) \bullet \log(p(i)) \qquad (14)$$

And then we use it as the loss for backpropagation, and use Adam to update the parameters in FABG.

Experiments

A. **Datasets**

Two datasets are applied: agnews and chnews.. Agnews is a collection of more than 1 million news articles. News articles have been gathered from more than 2000 news sources by ComeToMyHead in more than one year of activities. In this experiment, the topic classification dataset constructed by Zhang from this big dataset is used, which contains 4 categories (world, sports, business, sci/tech). Agnews is a balanced dataset. A dataset for news text classification with higher timeliness is proposed by collecting news in www.chinanews.com from January 1 to February 15, 2019, which is called chnews. This dataset is unbalanced. For this Chinese corpus, Jieba is used as word segmentation tool.

B. **Comparisons**

For comprehensive comparison, tthe results of some published papers is referred and implement some methods for text classification on the datasets used in this paper.

- Bag-of-words (BOW): BOW is a traditional method. BOW is constructed by selecting words from training set. The counts of each word is used as the features. The classifier is a multinomial logistic regression.
- CNN, Char-CNN and VDCNN: CNN for text classification was proposed by Kim. The features are extracted by using the convolution kernel size. Character-level CNN was proposed by Zhang. The very deep CNN model was proposed by Conneau.
- Bi-LSTM and Bi-GRU: It uses the Bi-LSTM and Bi-GRU to learn text semantics, then get the final representation of the text by avg-pooling.
- RCNN: RCNN was proposed by Lai. It contacts the outputs of the Bi-RNN and the word vectors itself as the input of max-pooling. Bi-GRU with attention: Since the datasets only contain short texts, we use the word encoder and word attention in HAN.
- FABG: FABG proposed in this paper with maxpooling layer and avg-pooling.

C. **Parameters**

Because the length of the input of FABG should be fixed, according to the length distribution of datasets in 4.1, the max length is 50 in agnews and 25 in chnews respectively. Texts longer than the max length are intercepted, and the shorter texts are padded by 0. 1 is the id of unknown words.

Conclusion

This paper proposes a full attention-based Bi-GRU neural network (FABG), which first uses the Bi-GRU to learn the semantic information of text, and then uses full attention layer to obtain the representation of each step by attending differently to the current and previous outputs of the Bi-GRU. Experiments show that FABG can achieve better results. Also, a Chinese dataset is built for news text classification by collecting news in www.chinanews.com. In recent years, the pre-trained models for word representation have made exciting improvements in NLP tasks, such as transformer, Bert, etc. These models can help obtain a better representation of the word, which provides a good idea that can further improve the performance of news text classification.