2019 IEEE 5th International Conference on Computer and Communications

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

Qinting Tang*, Jian li*, Jiayu Chen, Hengtong Lu, Yu Du, Kehan Yang
School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
e-mail: tangqinting@bupt.edu.cn, lijian@bupt.edu.cn

*Abstract*—**This paper proposes a novel approach for text classification by using attention mechanism. In recent works, several models based on deep learning with traditional attention mechanism mainly learn the weights of steps in the entire text. However, the information of each step is filtered by the encoder, and the same information has different effects on different steps. This paper proposes a full attention-based bidirectional GRU (Bi-GRU) neural network, which is called FABG. FABG uses a Bi-GRU to learn the semantic information of text, and uses full attention mechanism to learn the weights of previous and current outputs of the Bi-GRU at each step, which enables the representation of each step to obtain the important information and ignore the irrelevant information. Finally, through a pooling layer, we get the representation of the text. Thereby FABG can learn more information, which enhances the effect of text classification. Experiments on the English news dataset agnews and the Chinese news dataset chnews show that FABG achieve better performance than the baselines.**

***Keywords-attention; bidirectional GRU; text classification***

## I. 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) [1] is used to extract features from information of fixed-size windows, and recurrent neural network (RNN) based on long short-term memory (LSTM) [2] 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 [3] 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, we propose a bidirectional GRU (Bi-GRU) neural network based on full attention mechanism, 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.

Our main contributions in this work are: we propose a model with full attention mechanism, which can obtain richer key information. And experiments on two news datasets show that our model outperforms the baselines, which proves the effectiveness of our model. Also, we build a Chinese news dataset with higher timeliness by collecting news in www.chinanews.com from January to February 2019.

## II. RELATED WORK

Traditional methods of text classification are composed of three parts: feature engineering, feature extraction and machine learning algorithms. The most widely used method of feature engineering is Bag-of-word [4]. The most common method of feature extraction is to remove stop words [5]. And classifiers are the main method of machine learning algorithm, such as logistic regression [6], Bayes [7] and so on. However, these methods suffer from data sparse problem.

In order to overcome it, Hinton proposed deep learning. Deep learning can find the deep semantics in text, which makes it gain great success in NLP. Kim [8] used CNN to implement sentence-level text classification, and Zhang [9] put forward a character-level CNN. However, the single-layer CNN only obtains the information of fixed-size windows. So Conneau [10] raised a very deep CNN (VDCNN) and Johnson [11] raised deep pyramid CNN (DPCNN). But by deepening CNN, the ability of networks to capture the key information is still limited. Another popular neural network for text classification is RNN. Lai [12] came out with recurrent convolutional neural network (RCNN), which used bidirectional RNN to obtain context information and then selected features by a max-pooling layer. But in news text, the key information is useful, and RCNN can't utilize it. Attention mechanism, which is able to focus on the

978-1-7281-4743-7/19/$31.00 ©2019 IEEE

key information of the text, proves to be more useful for text classification. Yang [3] proposed HAN. By calculating the weights of words and sentences in the whole text, HAN can focus on important features like people, which strengthens the model's effect. The structure of HAN is very instructive to the work of this paper.

Inspired by HAN [3] and RCNN [12], this paper proposes FABG. Firstly, FABG learns the latent semantic information by the Bi-GRU. In full attention layer, FABG learns the influence of current and previous encoded outputs of the Bi-GRU on current step, then obtains a new representation of each step by weighted summing. Finally, pooling layer is used to get the features for classification.

The main difference between our work and previous works is the way to use attention mechanism. Previous works obtain the representation of the text by assigning attention weights to words and sentences and then get the label by that representation. However, the inputs of the attention layer have been filtered by the encoder, which means the key information could be missed. Also, the same information has different effects on different steps. So, we re-calculate the semantic representation of each step by adding the key information with attention mechanism. Then we get the final representation of the text by a pooling layer, finally get the label by that representation.

## III. 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 Fig. 1.

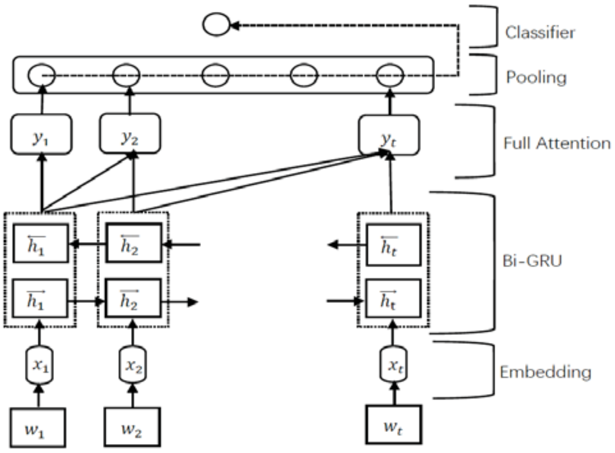

Figure 1. The architecture of FABG.

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 $R^{N\times d}$ 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 $x_t \in R^d$. The matrix of the input text is expressed as in

$$X = [x_1; x_2; \ldots; x_T] \in R^{T\times d} \tag{1}$$

### B. Encoder Based on Bi-GRU

Encoding the semantic information by Bi-GRU, FABG can learn the semantic dependence among words.

The GRU [13] is a variant of the LSTM [14]. 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

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

This is a linear interpolation between the previous state $h_{t-1}$ and the current candidate state $h_t$. 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

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

$r_t$ is the reset gate, $h_t$ can be calculated by

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

By the reset gate $r_t$, candidate state of step $t$ can obtain the information of input $x_t$ 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 $\overrightarrow{h_t}$ which reads the sentence from step $0$ to $t$ and the backward GRU $\overleftarrow{h_t}$ .

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

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

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

### C. *Full Attention*

In sentence-level text classification, the number of words in text is small. It's difficult to obtain more semantic information. As shown in Table 1, the key information of the text can reflect which category the text belongs to.

TABLE I. TABLE TYPE STYLES

| News | Key words | Category |
|---|---|---|
| **British** police have confirmed that the injuries at **Manchester** railway station are related to terrorism. | British;<br>Manchester | World |
| 117 environmental supervisors in **Shanxi**: two directors have been dismissed | Shanxi | Domestic |
| **Lei Wu** scored twice, **China's National football team** won the Philippines 3-0 and lead the way in advance. | Lei Wu;<br>China's National football team | Sport |
| **Banking Regulatory Commission**: 2019 will introduce measures to support the development of **microfinance** | Bank Regulatory Commission;<br>Microfinance | Finance |

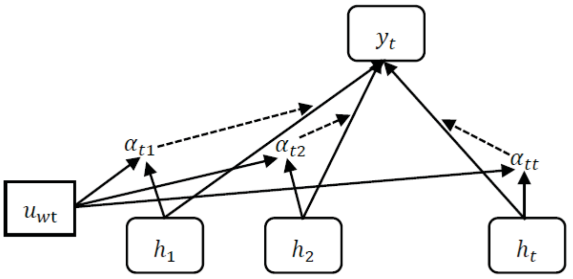

Figure 2. Example of a ONE-COLUMN figure caption.

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. The architecture of full attention at step $t$ is shown in Fig. 2

For step $t$, FABG contacts the outputs of the Bi-GRU to get the input of full attention layer

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

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 $y_t$ 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. In section 4, we will show the experimental results of these two different pooling layers.

### E. *Classifier*

We get the probability distribution by

$$p = soft\max(W_c Y + b_c) \tag{13}$$

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

$$\Lambda = -\sum_i q(i) \bullet \log(p(i)) \tag{14}$$

And then we use it as the loss for backpropagation, and use Adam [15] to update the parameters in FABG.

## IV. Experiments

### A. Datasets

We apply two datasets: agnews and chnews. The statistics is shown in Table 2. 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, we use the topic classification dataset constructed by Zhang [9] from this big dataset, which contains 4 categories (world, sports, business, sci/tech). Agnews is a balanced dataset. The distribution of length is shown in the Fig. 3. We propose a dataset for news text classification with higher timeliness by collecting news in www.chinanews.com from January 1 to February 15, 2019, which we call chnews. This dataset is unbalanced. The statistics is shown in Table 3. For this Chinese corpus, we use Jieba as word segmentation tool. The distribution of length is shown in the Fig. 4.

TABLE II. Statistics of Agnews and Chnews

| Dataset | Category | Train | Test | Ave-length | Max-length |
|---|---|---|---|---|---|
| Agnews | 4 | 120000 | 7600 | 31 | 173 |
| Chnews | 6 | 15000 | 6260 | 11 | 25 |

TABLE III. Statistics for Each Category of Chnews

| Category | Training | Testing | Total |
|---|---|---|---|
| Society | 5317 | 2204 | 7521 |
| Domestic | 3472 | 1441 | 4913 |
| World | 2420 | 998 | 3418 |
| Finance | 2283 | 961 | 3244 |
| Culture | 758 | 358 | 1116 |
| Sports | 750 | 298 | 1048 |

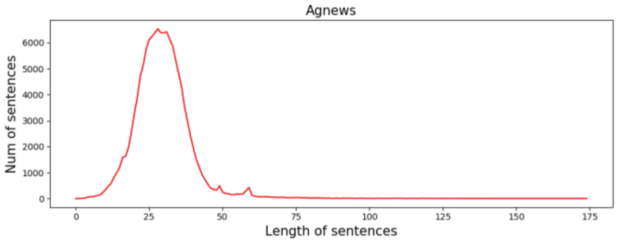

Figure 3. The distribution of length in agnews.

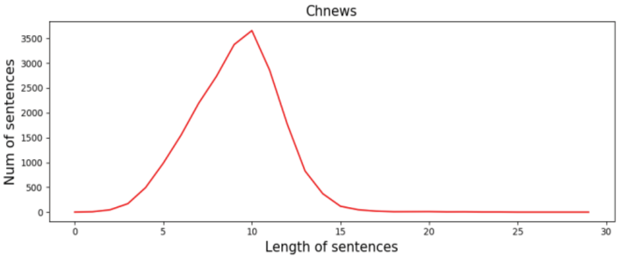

Figure 4. The distribution of length in chnews.

### B. Comparisons

For comprehensive comparison, we refer to the results of some published papers and implement some methods for text classification on the datasets used in this paper.

- **Bag-of-words (BOW)**: BOW is a traditional method used in [10]. BOW is constructed by selecting words from training set. We use the counts of each word as the features. The classifier is a multinomial logistic regression.
- **CNN, Char-CNN and VDCNN:** CNN for text classification was proposed by Kim [8]. The features are extracted by using the convolution kernel size [3,4,5]. Character-level CNN was proposed by Zhang [9]. The very deep CNN model was proposed by Conneau [10].
- **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 [12]. 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 [3].
- **FABG:** FABG proposed in this paper with max-pooling 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. The main hyperparameters and values are shown in Table 4.

TABLE IV. Hyperparameters and Values

| Parameters | Agnews | Chnews |
|---|---|---|
| Dimension of word vector | 100 | 100 |
| Hidden size | 100 | 50 |
| Attention size | 100 | 50 |
| Learning rate | 0.001 | 0.001 |
| Dropout | 0.5 | 0.5 |
| Batch size | 5 | 10 |

### D. Results and Analysis

The results of two datasets are shown in Table 5. The first part in Table 5 is BOW, which is the traditional method of text classification. The second part is models based on CNN. The third part is models based on RNN. The last part is models proposed in this paper. As we can see, FABG achieves the best results both on agnews and chnews.

By comparing the results of Bi-LSTM and Bi-GRU, we can see that the GRU is better than the LSTM. Because the GRU can effectively record and reset the context information in the text, we use the GRU as the component of the encoder in FAGB, which can make FABG work better. On the basis of learning the forward semantic information, the Bi-GRU adds the backward semantic information, which strengthens the ability of FABG to learn the context information.

TABLE V. RESULTS OF DIFFERENT MODELS

| Model | | Accuracy | |
|---|---|---|---|
| | | Agnews | Chnews |
| BOW [9] | | 88.8 | 44.66 |
| CNN [8] | | 89.47 | 62.60 |
| Char-CNN [9] | | 90.49 | - |
| VDCNN [10] | | 91.33 | - |
| RCNN [11] | | 90.45 | 63.02 |
| Bi-LSTM | | 90.21 | 62.70 |
| Bi-GRU | | 90.49 | 63.27 |
| Bi-GRU with attention | | 90.64 | 63.39 |
| FABG with | max-pooling | 91.50 | 65.48 |
| | avg-pooling | **91.79** | **65.65** |

Through the comparative analyses of the results of the Bi-GRU and the Bi-GRU with attention, we can observe that the latter performs better, which proves that attention mechanism can focus on the key information in news text classification. Although the Bi-GRU can obtain further context information, it can't pay attention to the key information which has a greater impact on new text classification. By learning the law in the text, attention mechanism can adjust the information obtained by the Bi-GRU, which can further enrich the semantic information. So, the Bi-GRU with attention mechanism can get better results. Based on this, FABG uses attention mechanism to make good use of the key information in text.

Compared to the Bi-GRU with attention, our FABG models achieve better performance. Traditional attention mechanism in the Bi-GRU can obtain the representation of text by attention weighted summing once. But the effect of key information on each step is different. Furthermore, the information is filtered by the Bi-GRU, which may result in missing the key information. Full attention mechanism in FABG can learn the weights of current and previous steps, and update the information at each step, which makes the vector of each step more informative.

By comparing the results of two pooling methods in FABG, we can observe that average pooling performs better than max pooling in FABG. The possible reason is that the representation of each step obtained by full attention layer has more information, avg-pooling layer is better by averaging all the features than by extracting the most important features directly through max-pooling layer.

Also, comparing with the traditional method, most recent deep learning methods can get better results. Especially in chnews that has more serious problem of sparse data, the effect of BOW is poor. For CNNs, since the single-layer CNN can only obtain context information of fixed-size windows, its result is poorer than VDCNN and DPCNN.

## V. 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, we build a Chinese dataset 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 [16], Bert [19], 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. In the following study, we will delve into how to use these pre-trained models to improve the effect of FABG on news text classification.

## REFERENCES

[1] Kalchbrenner N., Grefenstette E., Blunsom P.: A Convolutional Neural Network for Modelling Sentences. Eprint Arxiv (2014).

[2] Hochreiter S., Schmidhuber J.: Long short-term memory. Neural computation, 9(8):1735–1780 (1997).

[3] Yang Z., Yang D., Dyer C., et al.: Hierarchical Attention Networks for Document Classification. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2016).

[4] Wallach H. M.: Topic modeling: beyond bag-of-words. International Conference on Machine Learning. ACM (2006).

[5] Silva C., Ribeiro B.: The importance of stop word removal on recall values in text categorization. International Joint Conference on Neural Networks (2003).

[6] Ifrim G., Weikum G.: Fast logistic regression for text categorization with variable-length n-grams. Acm Sigkdd International Conference on Knowledge Discovery & Data Mining (2008).

[7] Genkin A., Lewis D. D., Madigan D.: Large-Scale Bayesian Logistic Regression for Text Categorization. Technometrics, 49(3):291-304 (2007).

[8] Kim Y.: Convolutional Neural Networks for Sentence Classification. Eprint Arxiv (2014).

[9] Zhang X., Zhao J., Yann L.: Character-level convolutional networks for text classification. In Adv. NIPS (2015).

[10] Conneau A., Schwenk H., Barrault, Loïc, et al.: Very Deep Convolutional Networks for natural language processing. arXiv prepint arXiv: 1606.01781 (2016).

[11] Johnson R., Zhang T.: Deep pyramid convolutional neural networks for text categorization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). volume 1, pages 562–570 (2017).

[12] Lai S., Xu L., Liu K., Zhao J.: Recurrent convolutional neural networks for text classification. In AAAI, 2267–2273 (2015).

[13] Cho K., Van Merrienboer B., Gulcehre C., et al.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Computer Science, (2014).

[14] Graves A.: Long Short-Term Memory. Supervised Sequence Labelling with Recurrent Neural Networks (2012).

[15] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. Computer Science (2014).

[16] Vaswani A., Shazeer N., Parmar N., et al.: Attention is all you need. in Conference on Neural Information Processing Systems (2017).

[17] Devlin J., Chang M. W., Lee K., et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805 (2018).