International Journal of Machine Learning and Cybernetics (2021) 12:1639–1647
https://doi.org/10.1007/s13042-020-01260-x

ORIGINAL ARTICLE

Hierarchical multi-attention networks for document classification

Yingren Huang[1,2] · Jiaojiao Chen[2] · Shaomin Zheng[2] · Yun Xue[2] · Xiaohui Hu[2]

Received: 14 August 2019 / Accepted: 9 December 2020 / Published online: 14 January 2021

Abstract
Research of document classification is ongoing to employ the attention based-deep learning algorithms and achieves impressive results. Owing to the complexity of the document, classical models, as well as single attention mechanism, fail to meet the demand of high-accuracy classification. This paper proposes a method that classifies the document via the hierarchical multi-attention networks, which describes the document from the word-sentence level and the sentence-document level. Further, different attention strategies are performed on different levels, which enables accurate assigning of the attention weight. Specifically, the soft attention mechanism is applied to the word-sentence level while the CNN-attention to the sentence-document level. Due to the distinctiveness of the model, the proposed method delivers the highest accuracy compared to other state-of-the-art methods. In addition, the attention weight visualization outcomes present the effectiveness of attention mechanism in distinguishing the importance.

Keywords Document classification · Hierarchical network · Bi-GRU · Attention mechanism

1 Introduction

During the past decades, the growth of the internet has sparked an explosion in the rate at which text data is produced and published. Large amount of the information available is stored as documents; therefore, the organization of that information has become a complex and vital task. Document classification, which is one of the focus issues in Natural language processing (NLP), plays a crucial role in the ability of sorting, directing, classifying and providing the proper information in a timely and correct manner [1]. Generally, document classification, contains spam filtering, email categorization, information retrieval, sentiment analysis, etc. [2, 3]. As such, it is essential that we develop methods for dealing with large collections of documents. Considering the significance of document classification, numerous state-of-the-art algorithms are utilized and notable progresses are achieved [4]. Now that document classification is a supervised learning approach, each document's category is learned from a set of texts with predefined labels [5, 6]. Distinctively, the working performance of classification varies based on the techniques used to develop it [7].

More recently, the flourishing of the machine learning techniques provides researchers with more opportunities to resolve the classification issue. The supervised learning algorithms, such as decision tree and support vector machines (SVM), are applied to document classification and their working performance are evaluated [8–10]. Recently, deep-learning models have attracted a great deal of interest in document classification. On the task of model establishing, both convolutional neural networks (CNN) and recurrent neural networks (RNN) are specifically pronounced. Yoon Kim applies CNN to sentence classification [11]. Zhao et al. proposed a fusion ELMo and Multi-Scale Convolutional Neural Network (MSCNN) as classifiers based on the scale features of different scales [12]. Similarly, Kalchbrenner et al. introduces the dynamic

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13042-020-01260-x.

✉ Yun Xue
xueyun@scnu.edu.cn

Yingren Huang
yrhuang@m.scnu.edu.cn

[1] Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou, Guangdong, China

[2] Guangdong Provincial Key Laboratory of Quantum Engineering and Quantum Materials, School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou 510006, China

convolutional neural network (DCNN) to semantic modelling of sentences, in line with a global pooling operation over linear sequences [20]. On the other hand, RNN is capable of retaining the memory of all the previous text in a sequence of hidden states [13]. A discourse model on the foundation of a novel RNN is established to obtain a sound outcome in a dialogue classification experiment [22].Seeing that remarkable progresses are made in CNN and RNN models, improving strategies for deep learning algorithms are put forward. Specifically, attention mechanism is highlighted in the ongoing research. The attention mechanism is originally proposed referring to human visual focus to acquire information and achieves an appealing performance in image recognition [14]. In 2014, Bahdanau et al. introduces attention mechanism was integrated into RNN model for machine translating and outperforms traditional statistical machine translation [15]. Likewise, Du et al. exploit the possibilities of attention mechanism in text classification. Encouragingly, the proposed model is capable of extracting very salient parts from sentences [16].

The use of attention mechanism in RNN model for document classification is, however, still limited, primarily because of its impossible to deal with the hierarchical structure information. Further, the RNN is employed for merely exploiting the historical context, which fails to deal with the local and succeeding contexts for the position in a sequence [17]. That is, reliably retaining long-range information is a well-documented weakness of RNN networks. As a result, hierarchical architecture seems to be a sound solution to obtain a better performance [18, 19]. In addition, according to [20], the structure property of the document cannot be fully adopted with merely single pattern of attention mechanism. It is intuitive that not all parts within the document are equally influential on delivering the main idea. As reported in [19], with different attention applied to specific categories, the classification accuracy can be improved to a large extent. For these reasons, attention mechanism in RNN model is generally taken as an alternative for document classification.

In this research, we concentrate on the distinctiveness of the document to identify the connections between words and sentences as well as sentences to the document. Thus, the deep learning network is applied to the hierarchical structure of the document. Also, different attention strategies are integrated according to the different levels for attention weights determination. In this way, a high-accuracy document classification model can be designed and deployed. Our contributions are summarized below:

1.A specific architecture using bidirectional gated recurrent unit (Bi-GRU) network for document classification is established, which, based on the relation within the document, contains two-stages of hierarchical structure: word-to-sentence level and sentence-to-document level.

2.Aiming to identify the significances of different words and different sentences, two distinguishing attention mechanisms are employed on each level for modeling. In this way, we target at computing more precise attention weights of different parts.

3.The proposed framework takes advantages of the deep averaging networks (DAN) and deep vector averaging (DVA) for information capturing. Empirical results demonstrated that our model achieves an even higher accuracy comparing to the state-of-the-arts.

This paper will introduce the background knowledge in Sect. 2, illustrate the architecture of the hierarchical multi-Attention networks (HMAN) in Sect. 3, shows the results achieved in experiment and the analysis of the model in Sect. 4, and finally presents the research findings and future expectation in Sect. 5.

2 Background knowledge

This section will introduce the basic knowledge related to RNN based classifiers, so as to facilitate the description of subsequent model establishment.

2.1 Bidirectional-GRU network

GRU was initially proposed by Cho et al. to make each recurrent unit to adaptively capture dependencies of different time scales [21]. And GRU has been proven to be effective in a variety of applications [22]. A standard architecture of GRU is shown in Fig. 1.

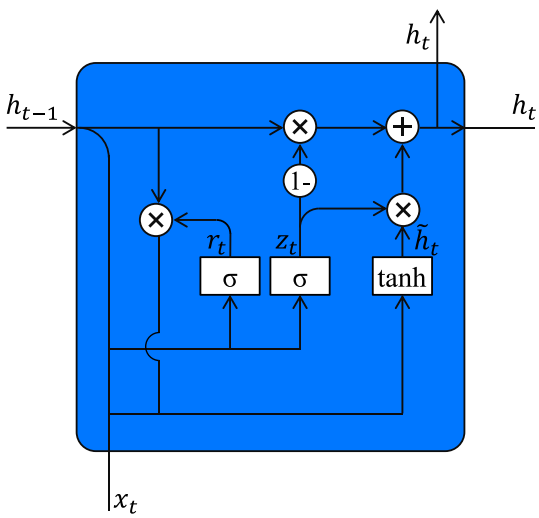

Fig. 1 Structure of the GRU

A GRU unit employs the gating mechanism for state monitoring. Explicitly, there are two gates in one GRU unit, namely reset gate r and update gate z. For a specific time step t, we have

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1} + b_r\right) \tag{1}$$

$$z_t = \sigma\left(W_z x_t + U_z h_{t-1} + b_z\right) \tag{2}$$

$$\tilde{h}_t = tanh\left(W_h x_t + U_h\left(h_{t-1} \odot r_t\right) + b_h\right) \tag{3}$$

$$h_t = z_t \odot \tilde{h}_t + \left(1 - z_t\right) \odot h_{t-1} \tag{4}$$

where W and U stand for the weight matrices that can be trained and b is the bias vector. The symbol $\odot$ is an element-wise multiplication while σ indicates a logistic sigmoid function with the output value in [0, 1]. Besides, for x is the current input, we also take h_{t-1} as the state of previous step. The output vector h_t is made up of the candidate output $\tilde{h}_t$, which is computed considering the reset gate in Eq. (3). Hereafter, the update gate determines the previous information to be removed and new information to be added, as given in Eq. (4). The smaller value of reset gate is, the more previous memory will be neglected; The larger value of update gate is, the more previous information will be brought [23].

Similarly, as long as two GRU units of reverse timing are connected to one output, a Bidirectional-GRU unit related to sequence context is obtained, which is delivered as:

$$\begin{cases} \overrightarrow{h_t} = \overrightarrow{GRU}\left(x_t\right), t \in [1, T] \\ \overleftarrow{h_t} = \overleftarrow{GRU}\left(x_t\right), t \in [T, 1] \\ h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] \end{cases} \tag{5}$$

2.2 Deep averaging networks (DAN)

The DAN was originally designed for feeding an unweighted average of word vectors through multiple hidden layers before text classification [24]. An example is illustrated in Fig. 2. The basic working principle is concluded as follows. The vector average of the embeddings related to an input sequence of tokens is computed. Thereby, the obtained average is passed through the nonlinear layers for semantic information extraction. The outcome is given on the final layer's representation. On the other hand, the model robustness is improved via the novel dropout-inspired regularizer. During the training process, some of the tokens' embeddings are randomly dropped beforehand. In this way, a DAN model can be defined as:

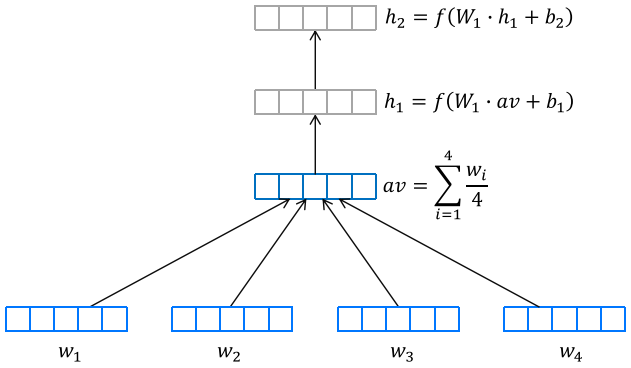

Fig. 2 An example of DAN

$$DAN(X) = MLP(Average(Dropout(X))) \tag{6}$$

where MLP refers to the Multilayer perceptron computing basis in deep learning.

2.3 Deep vector averaging (DVA)

As shown in Fig. 3, the DVA model is composed of an ordered RNN and an unordered DAN, which can be expressed as:

$$DVA(X) = [RNN(X), DAN(X)] \tag{7}$$

One of the key facts can be observed from RNN is that the information as well as the $n-gram$ within a long sequence document cannot be well-retained [25, 26]. Nevertheless, the use of DAN addresses this issue by integrating the observed effectiveness of depth with the unreasonable effectiveness of unordered representations. What's more, the DVA can not only be applied to RNN model, but also to the improved model based on RNN architectures, such as Long Short-Term Memory (LSTM) and GRU. Accordingly, the revised model is able to capture the uni-gram and bi-gram occurrences over long sequences. For more detailed analysis see for example [4].

3 Approach

In this section, we present the details of the model HMAN, which consists of two layers of network corresponding to the word-to-sentence and the sentence-to-document analysis, respectively.

We establish a classification model which aims at processing the document at two different levels, which are word-sentence level and sentence-document level (Fig. 4). The soft-attention mechanism and the CNN-attention mechanism are performed on the word-sentence level and

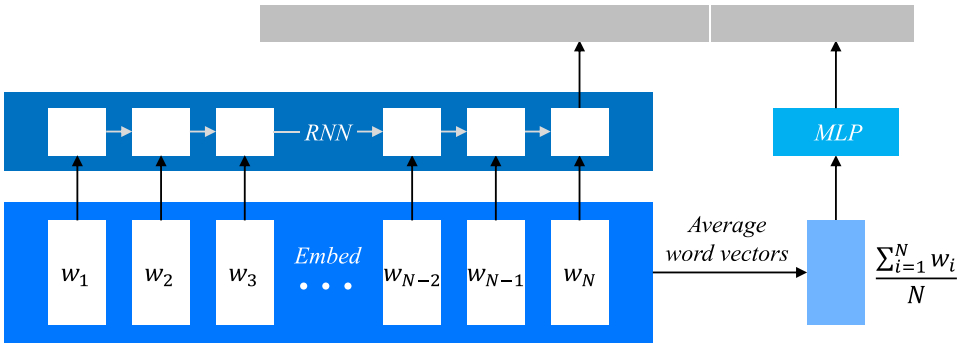

Fig. 3 Structure of the DVA model

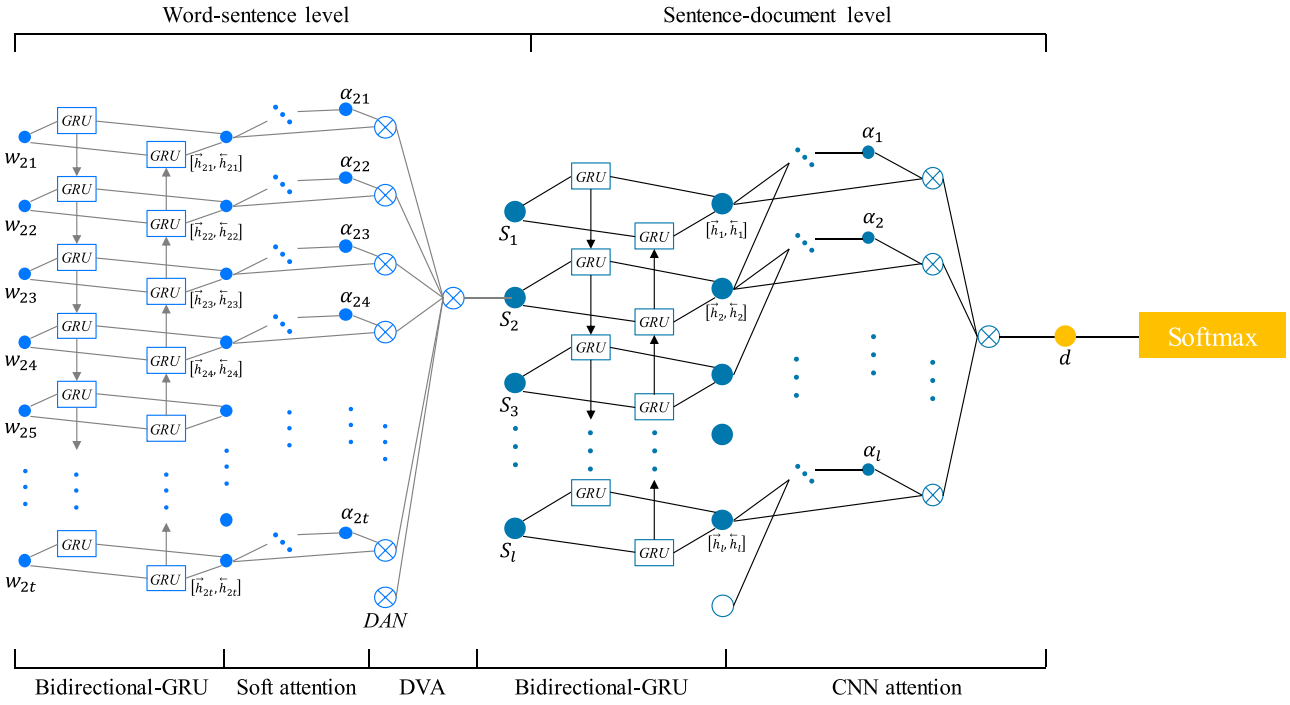

Fig. 4 Architecture of HMAN

sentence-document level, respectively. Distinctively, in a single sentence, the words are generally arranged non-sequentially. Seeing that soft-attention is able to adaptive to extract the semantic within a word collection, i.e. insensitive to the word order, we thus apply it to the word-sentence level. On the other hand, the N-gram based attention networks [16], such as CNN attention-mechanism, are deemed best to deal with the coherence among sentences. In terms of the sentence-document level, CNN-attention mechanism, which can highlight not only the content but also the local relation, is taken as a better alternative.

This idea is to transform the document into distinguished components that evaluates how probable it is to take a specific network in accordance to the target component, and then use the network to seed a classification algorithm. The construction of the model is based on basic deep learning network, with the attention mechanism integrated. The primary stages of the proposed model can be conveyed as follows:

1.Encoding every single word in the sentence using Bi-GRU model, thus the context information among words is collected.

2.For each sentence, the attention weights of words, which represent the importance, is assigned via the Soft attention strategy [20].

3.The attention vector of one sentence is therefore computed by weighting each word in this sentence.

4.Since DAN is employed for semantic processing, the DAN outcome together with the sentence attention are

integrated using DVA network to determine the vector for representing the sentence.

5.Similarly, the sentences in the document are encoded in the same way as the words.

6.The attention weight that attaches to each sentence is determined via the CNN attention strategy [16].

7.The document is characterized by a vector based on weighting each sentence in this document.

8.The final result of document classification is given through a softmax classifier.

3.1 Word-sentence level

Let's concentrate now on the structure of the document. Suppose that there are L sentences in the document. We shall also define the ith sentence contains T_i words and x_{it} is the tth one, which can be vectorized as w_{it}.

As pointed out beforehand, the words are encoded by Bi-GRU network. Thus, we have

$$\begin{cases} \overrightarrow{h_{it}} = \overrightarrow{GRU}(w_{it}), t \in [1, T_i] \\ \overleftarrow{h_{it}} = \overleftarrow{GRU}(w_{it}), t \in [T_i, 1] \\ h_{it} = \left[\overrightarrow{h_{it}}, \overleftarrow{h_{it}}\right] \end{cases} \tag{8}$$

where h_{it} is the Bi-GRU output, consisting of $\overrightarrow{h_{it}}$ and $\overleftarrow{h_{it}}$, with context information included.

Considering that each word may be of different importance for the sentence, the Soft attention is employed for assigning of the weight according to the context information. On this occasion, h_{it} is taken as the input of one-layer MLP and thus a u_{it} is obtained:

$$u_{it} = tanh\left(W_w h_{it} + b_w\right) \tag{9}$$

With the application of softmax model, the normalized attention weight α_{it}, which indicates the contribution to the sentence, can be written as

$$\alpha_{it} = exp\left(u_{it}\right) / \sum_t exp\left(u_{it}\right) \tag{10}$$

In relation to the semantic delivery, we can calculate the sentence attention vector from the word attention. According to Eq. (10), we have

$$\underset{s_i}{\sim} = \sum_{t=1}^{T_i} \alpha_{it} h_{it} \tag{11}$$

where $\underset{s_i}{\sim}$ stands for the sentence attention vector acquired from weighted summation. Note that $\underset{s_i}{\sim}$ is also sent to the DVA network, when the other component is the semantic information coming from the DAN. Based on the working principle of DAN, we define now its outcome as

$$DAN\left(\sum_{t=1}^{T_i} w_{it}\right) = MLP\left(Average\left(\sum_{t=1}^{T_i} Dropout\left(w_{it}\right)\right)\right) \tag{12}$$

A complete sentence vector s_i is obtained via the integration of both parts

$$s_i = \left[\underset{s_i}{\sim}, DAN\left(\sum_{t=1}^{T_i} w_{it}\right)\right] \tag{13}$$

3.2 Sentence-document level

Likewise, the Bi-GRU can also be utilized to encode the sentences, i.e.

$$\begin{cases} \overrightarrow{h_i} = \overrightarrow{GRU}(s_i), i \in [1, T_i] \\ \overleftarrow{h_i} = \overleftarrow{GRU}(s_i), i \in [T_i, 1] \\ h_i = \left[\overrightarrow{h_i}, \overleftarrow{h_i}\right] \end{cases} \tag{14}$$

where h_i is the output of Bi-GRU for sentence encoding. In line with the current research findings, the window for sentence processing slides over several adjacent sentences at one time, which indicates more relations exist among the neighboring [17]. Consequently, on the foundation of local correlation within adjacent sentences, we set the CNN attention strategy to determine the attention weight of each sentence (Fig. 5). Now that the CNN algorithm originates from the biological vision principle, the local features of inputs can be extracted via different tools, such as multi-network, convolution and down-sampling. Further, the output of Bi-GRU, without loss of context information, is taken as the input of CNN. In this way, both the context and the local correlation are kept during processing.

As shown in Fig. 5, let

$$h_{i:(i+k-1)} = \left[h_i; h_{i+1}; \cdots; h_{i+k-1}\right] \quad i \in [1, L] \tag{15}$$

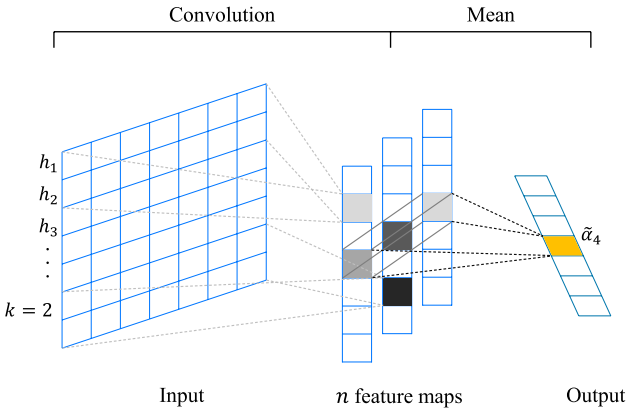

Fig. 5 An example of CNN attention model

be the status of hidden layers where k represents the window of CNN and $k = 2$. Specifically, if $i + k - 1 > L$, we use zero vector for slide window supplementary. In this stage, n different convolutional filters are used in a convolutional layer. For the i^{th} input sentence h_i, the weight vector $\overset{\sim}{\alpha}_{ji}$ by using the convolutional filter

$$\overset{\sim}{\alpha}_{ji} = \sigma\left(f_j \circ h_{i:(i+k-1)} + b_j\right) \; j \in [1, n] \tag{16}$$

where $f_j \in R_{kd}$ stands for the jth convolutional filter. With all the convolutional filters applied, we get n different outcomes of that sentence. For each sentence, the convolutional result is calculated by averaging, which is

$$\overset{\sim}{\alpha}_i = \sum\nolimits_j^n \overset{\sim}{\alpha}_{ji} / n \tag{17}$$

As long as the outputs of all sentences from the CNN model are recorded, the attention weight of every single sentence can be presented by normalization, which are computationally efficient to work with.

$$\alpha_i = \overset{\sim}{\alpha}_i / \sum\nolimits_{i=1}^{L} \overset{\sim}{\alpha}_i \; i \in [1, L] \tag{18}$$

The expression of the document is in fact the weighted summation of all sentences in it.

$$d = \sum\nolimits_{i=1}^{L} \alpha_i h_i \tag{19}$$

where d indicates the vector of the target document.

Finally, the document representation is sent to a softmax classifier for labeling where the conditional probabilities over the class space are produced. The class label is given by

$$o = softmax(Wd + b) \tag{20}$$

4 Experiments

In this section, we illustrate our experiments together with the results on real comments. The dataset for our model evaluation is provided. We compare the HMAN with several other models by obtaining a comprehensive five-classes outcomes.

4.1 Dataset

The user comments employed in this research come from three network sites. Concretely, Yelp reviews are extracted from Yelp, which is one of the most famous review sites in the US. There are over 4.7 million consumers' reviews with the rating from 1 to 5. To facilitate the computation, we develop two comprehensive datasets, namely Yelp1 and Yelp2, by randomly picking reviews from the network where Yelp1 and Yelp2 contains 1.99 million and 1.98 million reviews, respectively. Amazon Fine Food Reviews and Amazon Mobile Phones Reviews are exactly the reviews of food and phones on Amazon.com. Both of the datasets has 110,000 reviews while all are applied to the model studying. Reviews from Amazon are with the rank of 1 to 5, which are similar to those from Yelp. The whole data is randomly split into training, validation and testing according to the proportion of 80%, 10% and 10%. Hyperparameters are finetuned in line with the validation outcomes. More details about the datasets in this experiment is shown in Table 1.

4.2 Experimental setup

The collected reviews are preprocessed by using the NLTK (Natural Language Toolkit) tool, which is a suite of open source program modules based on Python. Formally, both sentence segmentation, and word segmentation are applied to all documents. Meanwhile, all documents are sent to procedures of punctuation removing and upper-lower case conversion. In this research, each employed word vector is of 300 dimension and is established from a randomized way. The model is established and trained based on TensorFlow. The main parameters for model training are fixed as shown in Table 2.

4.3 Baselines

Aiming at evaluating the working performance of the proposed model, other state-of-the-art methods are taken into the experiment. We now describe the comparison methods as follows:

4.3.1 Bi-GRU

This method uses the bidirectional gated recurrent unit framework which encodes and decodes the document materials [27].

4.3.2 Bi-GRU + attention

Basic single-level Gated Recurrent Unit with the attention mechanism added. The attention weight is assigned to different part within the document.

4.3.3 CRAN

One-layer network based on LSTM proposed in [16] integrated with the CNN attention.

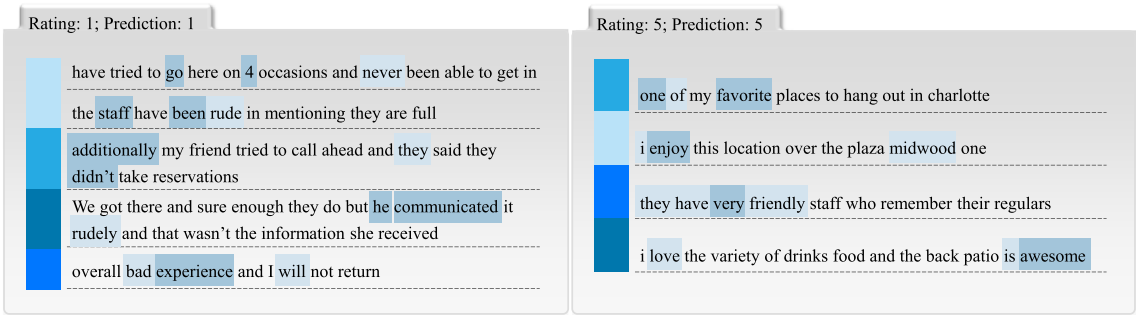

Fig. 6 Attention weight of documents

Table 1 Data statistics

Dataset	Classes	Documents	Average#s	Max#s	Average#w	Max#w
Yelp1	5	1,990,636	8.289	171	117.878	1090
Yelp2	5	1,894,817	8.279	151	117.751	1090
Food reviews	5	110,000	2.959	222	42.028	5668
Phone reviews	5	110,000	4.980	346	86.168	3659

#s represents the number of sentences (average and maximum per document), #w represents the number of words (average and maximum per document)

Table 2 Working parameters for HMAN model

Parameter	Value
Learning rate	0.001
Batch sizes	256
Truncated sentences	30
Truncated words	30
Hidden units	100
Convolutional windows	5
Convolutional filters	256
Dropout rate	0.5
Epoch	200

Table 3 Average accuracy of different methods

Methods	Yelp1	Yelp2	Fine food reviews	Mobile phones reviews
Bi-GRU	70.7	70.6	81.6	81.2
Bi-GRU + Attention	71.6	71.8	81.9	82.2
CRAN	71.4	71.4	81.7	82.7
HCRAN	73.2	73.2	82.5	81.8
HCRAN + DVA	73.4	73.3	82.5	83.0
HAN	73.4	73.3	82.4	82.9
HAN + DVA	73.4	73.4	82.8	83.0
HMAN	73.4	73.4	82.6	83.5
HMAN-no DVA	73.4	73.4	82.8	82.3

4.3.4 HCRAN

Integration of LSTM and hierarchical attention mechanism, which establishes two-level architecture by using CNN attention mechanism [28].

4.3.5 HAN

Hierarchical Attention Network based on RNN developed by Yang et al. [20]. A two-level network is designed with soft-attention integrated.

Consequently, we present the average accuracy of these methods on the aforementioned datasets.

5 Result

As mentioned before, the experiment involves the dataset comprising four review groups. Multiclass classification experiments are performed to evaluate the working accuracy. Table 3 summarizes the experimental outcomes for the aforementioned methods, respectively.

To further investigate the impact of different components, we also integrate two baseline models with DVA, which are HCRAN + DVA and HAN + DVA. In this way, the significance of DVA in long-distance related information processing is determined. Likewise, a HMAN without DVA (HMAN-no DVA) is also performed. The ablation

of DVA in HMAN results in dealing with only the text sequence information.

6 Discussion

Clearly, there is a considerable gap between the accuracy of "Mobile phone reviews" and "Fine food reviews" and the other two. A possible explanation is that more professional words are used to describe the feelings about the specific categories, which are distinctive from others and are easy to identify. With the attention mechanism integrated, a considerable improvement of the working accuracy is brought, i.e. the accuracy of Bi-GRU + attention and CRAN is much higher than that of Bi-GRU. We shall thus say that the attention mechanism can be employed for precisely figuring out the differences of importance in the document, which results in upgrading the classification accuracy by acquiring more effective information. Similarly, a more than 2% improvement in accuracy is obtained in HMAN compared to that of Bi-GRU.

Furthermore, as for the network integrated with attention mechanism, i.e. Bi-GRU + attention and CRAN, the accuracy of HMAN on each dataset is even higher, which indicates the hierarchical structure outperforms the single-layer network in classification. The average accuracy of HMAN is 1.65% and 1.67% higher in comparison to Bi-GRU + attention and CRAN, respectively. The idea of classifying the document at different levels, which is theoretically in a more efficient manner of human processing, clearly benefits the working performance. The forming of DVA, on the basic of algorithms combination, addresses the long-distance information delivery issue via the hidden layer of Bi-GRU.

Regarding the hierarchical attention networks, results show that our model is a better alternative to HAN and HCRAN. The average accuracy of HMAN is about 0.4% higher than that of HAN while 0.6% higher than that of HCRAN. On the task of identifying the sentiment of the document, our model delicately applies distinguishing attention mechanisms to assign the attention weights more precisely. Moreover, we observe that the withdrawal of DVA in HMAN cause a 1.2% drop on the accuracy of dataset 'Mobile phone reviews'. Nevertheless, the distinction of DVA contribution on the other three datasets is relatively small. In addition, with the integration of DVA, the accuracy of HAN + DVA and HCRAN + DVA is improved as well. Thus, considerably more informative representations can be obtained due to the employment of DVA. It is reasonable to expect better model formation and thus better working performance, as it is the case.

6.1 Visualization of attention

Considering the capability of attention mechanism, it is advisable to visualize the attention weight presentation. In this case, we illustrate the distinctive importance of words and sentences identified by the proposed model. As shown in Fig. 6, sentences and words in darker color are have greater weight, and vice versa. For instance, in the last sentence, HMAN model initially assigns the importance to the verbs, i.e. "love" and "is", to characterize the main opinion of the sentence. Also, the word "awesome" is given a higher attention value due to its carrying a strong sentiment, which makes the sentence of great importance as well. In this way, the outcome of the document can be predicted.

7 Conclusion

In this paper, the HMAN for document classification is described. The deployment is obtained via the integrated of deep learning model and the attention mechanism. The model is established via the DVA algorithm, which is applied to facilitate the processing by using the semantic information. We then construct the model, considering the word-sentence level and the sentence-document level, to employ different attention strategies to assign the attention weight effectively. Experiments are carried out on typical comment datasets with the experimental outcomes carefully analyzed. Compared to some state-of-the-art methods, the results validate the effectiveness of HMAN and demonstrate the high accuracy in document classification.

Future work will address the optimization of the proposed model. As long as the complex structure will result in the issues like slow convergence, other methods can be integrated to finetune the current model to further improve the working performance of classification.

Acknowledgements This work was supported by the National Statistical Science Research Project of China under Grant No. 2016LY98, the Science and Technology Department of Guangdong Province in China under Grant Nos. 2016A010101020, 2016A010101021 and 2016A010101022, the Characteristic Innovation Projects of Guangdong Colleges and Universities (Nos. 2018KTSCX049 and 2018GKTSCX069), the Bidding Project of Laboratory of Language Engineering and Computing of Guangdong University of Foreign Studies (No. LEC2019ZBKT005).

References

1. Chambers A (2013) Statistical models for text classification and clustering: applications and analysis. Dissertation. University of California, Irvine
2. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41:391–407
3. Pang B, Lee LJF (2008) Retrieval TiI. Opin Min Sentiment Anal 2:1–135
4. Longpre S, Pradhan S, Xiong C, Socher RJapa (2016) A way out of the odyssey: analyzing and combining recent insights for LSTMs
5. Sarioglu ES (2014) Effective classification of clinical reports: natural language processing-based and topic modeling-based approaches. The George Washington University
6. Hassan A, Mahmood A (2018) Convolutional recurrent deep learning model for sentence classification. IEEE Access 6:13949–13957. https://doi.org/10.1109/ACCESS.2018.2814818
7. Core DB (2012) Applications of text classification to enterprise support documents. UC Santa Cruz
8. Silva C, Lotric U, Ribeiro B (2010) Dobnikar AJIToS, Man, cybernetics PC. Distributed text classification with an ensemble kernel-based learning approach. IEEE Trans Syst Man Cybern 40:287–297
9. Nii M, Ando S, Takahashi Y, Uchinuno A, Sakashita R (2007) Nursing-care freestyle text classification using support vector machines. In: IEEE international conference on granular computing, 2007, GRC 2007. IEEE, New York, pp 665–665
10. Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. In: ICML, 1997, pp 412–420
11. Kim Y (2014) Convolutional neural networks for sentence classification arXiv preprint https://arxiv.org/abs/14085882
12. Zhao Y, Zhang J, Li Y et al (2019) Sentiment analysis using embedding from language model and multi-scale convolutional neural network. Comput Appl 40(3):651–657
13. Funahashi K-I, Nakamura Y (1993) Approximation of dynamical systems by continuous time recurrent neural networks. Neural Netw 6:801–806
14. Mnih V, Heess N, Graves A (2014) Recurrent models of visual attention. In: Advances in neural information processing systems, pp 2204–2212
15. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint https://arxiv.org/abs/14090473
16. Du J, Gui L, Xu R, He YA (2017) Convolutional attention model for text classification. In: National CCF conference on natural language processing and Chinese computing, 2017. Springer, New York, pp 183–195
17. Zhang Y, Er MJ, Venkatesan R, Wang N, Pratama M (2016) Sentiment classification using comprehensive attention recurrent models. In: 2016 international joint conference on neural networks (IJCNN). IEEE, New York, pp 1562–1569
18. Remy JB, Tixier AJP, Vazirgiannis M (2019) Bidirectional context-aware hierarchical attention network for document understanding. arXiv preprint https://arxiv.org/abs/1908.06006
19. Shi M, Liu J (2018) Functional and contextual attention-based LSTM for service recommendation in Mashup creation. IEEE Trans Parallel Distrib Syst 30(5):1077–1090
20. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, pp 1480–1489
21. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint https://arxiv.org/abs/14123555
22. Xu C, Shen J, Du X, Zhang F (2018) An intrusion detection system using a deep neural network with gated recurrent units. IEEE Access 6:48697–48707. https://doi.org/10.1109/ACCESS.2018.2867564
23. Song Y (2018) Stock trend prediction: based on machine learning methods. UCLA
24. Iyyer M, Manjunatha V, Boyd-Graber J, Daumé III H (2015) Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: long papers), pp 1681–1691
25. Karpathy A, Johnson J, Fei-Fei L (2015) Visualizing and understanding recurrent networks arXiv preprint https://arxiv.org/abs/150602078
26. Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies, vol 1. Association for Computational Linguistics, pp 142–150
27. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2016) Learning phrase representations using RNN encoder-decoder for statistical machine translation, arXiv preprint https://arxiv.org/abs/1406.1078.
28. Du J, Gui L, He Y et al (2019) Convolution-based neural attention with applications to sentiment classification. IEEE Access 7:27983–27992