Improved Aspect-level Sentiment Analysis Method based on Multi-Head Attention Mechanism

Kaixuan Yu\textsuperscript{1}, Yachao Li\textsuperscript{1} and Dongsheng Zhang\textsuperscript{1}\textsuperscript{*}

\textsuperscript{1}School of Software, Henan University, Kaifeng, Henan, 475000, China
\textsuperscript{*}zds@vip.henu.edu.cn

Abstract: The purpose of the aspect-level sentiment analysis task is to analyze the sentiment orientation expressed by different aspects in the text. It has more fine-grained sentiment evaluation objects and is more in line with actual needs. Therefore, it has received extensive attention in recent years. At present, the model that combines the attention mechanism with the Recurrent Neural Network and its variants has gradually become the main method to solve aspect level sentiment analysis tasks. However, this type of method is limited by the deficiencies of the RNN itself, the training time is too long and the dependence between words decreases with the increase of distance. Even if the attention mechanism is added, the above problems will still exist. Therefore, this paper introduces the multi-head attention mechanism into the GCAE model (Gated Convolutional network with Aspect Embedding), and then proposes the GCAE-MHA model (Improved GCAE model based on Multi-Head Attention mechanism). This model models the context and specific aspects of the text sequence at the same time, and learns the interactive relationship between them. And finally uses the two together as the basis for sentiment classification. At the same time, in order to make up for the problem that the convolutional neural network can only extract the dependencies between local words, the GCAE-MHA model also uses the Dilated Convolutional Neural Network to replace it, which can extract the semantic information between long-distance words, and by setting different dilation rate to obtain a richer semantic feature. Finally, experiments are carried out on the SemEval2014 dataset and Twitter dataset. The experimental results show that the GCAE-MHA model can effectively improve the effect of aspect level sentiment analysis while ensuring the simplicity of the model network structure.

1. Introduction
With the development of Internet technology, more and more social media emerge in an endless stream. While these applications bring great convenience to people’s lives, they also provide convenient ways for people to express their opinions on certain things. The content of personal sentiment contains huge commercial value and social value, so sentiment analysis tasks have been very active in natural language processing tasks.

According to the different mining granularity, sentiment analysis tasks can be divided into three levels: document-level sentiment analysis, sentence-level sentiment analysis, and aspect-level sentiment analysis\textsuperscript{1}. Among them, document-level and sentence-level sentiment analysis is to make a judgment on the overall sentiment orientation of the text(such as positive, negative, and neutral). Although this kind of coarse-grained sentiment analysis method is suitable for solving some problems, such as analyzing the overall sentiment polarity of product reviews in order to calculate the favorable rate. However, as people's demand for understanding the details of viewpoint mining objects increases, its limitations are getting bigger and bigger. In fact, when people express their opinions, they may express...
different emotional tendencies in different aspects of the same thing, and only taking an overall emotional state cannot fully express the true emotions it contains. For example, "The food in this restaurant is delicious, but the environment is terrible." It can be seen that opinion holders express completely opposite emotional tendencies on the two aspects of "food" and "environment." Sentiment analysis technology, whether it is to predict the text as positive or negative, is inappropriate. It is in the context of this demand that the aspect-level sentiment analysis task came into being. Aspect-level sentiment analysis was formally proposed by Thet in 2010[2], whose task is to extract and summarize the emotional polarity expressed in different aspects of the text, needs to consider both the contextual information and specific aspects of the text, which can more effectively help people understand the emotional tendencies of different aspects, and has more important practical significance.

In recent years, deep learning technology has achieved rapid development, and has achieved excellent results in various fields, which has gradually become the mainstream method in the field of aspect-level sentiment analysis. RNN and its variants adopt a sequential calculation method, have a certain "memory ability", and are very effective in processing sequential data. So they are widely used in aspect-level sentiment analysis tasks. Tang[3] proposed a method TD-LSTM based on the Long Short-Term Memory network (LSTM), which splits the sentence into two parts according to the location of the word in a particular aspect. And uses two LSTM networks obtain the left and right hidden layer outputs to obtain the classification results. Attention Mechanism originated from the field of computer vision[4], and its purpose is to make the network model pay close attention to those parts that need attention while ignoring those parts that do not need to be paid attention to. After that, the attention mechanism has been widely used in various tasks in the field of natural language processing[5,6]. For aspect-level sentiment analysis tasks, the attention mechanism can effectively focus on important aspects related to the context, so many scholars combine the attention mechanism with neural network models to handle aspect-level sentiment analysis tasks. The ATAE-LSTM model[7] is a typical method based on LSTM network and attention mechanism. This method connects the aspect embedding with its context representation, and then uses LSTM and attention mechanism to generate the final representation, which verifies the effectiveness of the attention mechanism. Aiming at the problem of less interaction between specific aspects and context, Ma[8] proposed an interactive attention network model IAN, which can simultaneously focus on important context information and aspect information in a sentence, verifying that the text context and aspect sequence the importance of interaction. Huang[9] proposed the AOA-LSTM model, which uses the LSTM network to simultaneously model the context and aspect sequence, and then uses the AOA (Attention-over-Attention) module[10] to learn the interaction relationships between the context and the aspect, once again verify the importance of context and aspect interaction. The above research work has explored aspect-level sentiment analysis tasks from different angles, and has achieved certain results. But the above methods will be limited by RNN and its variants to varying degrees. Specifically, RNN and its variants adopt a sequential calculation method, which is difficult to achieve parallelization and expensive training costs. In addition, as the distance increases, the dependency between words will weaken or even disappear, making it difficult to capture the complete structure information of the sentence, which will undoubtedly have a certain impact on the final sentiment classification effect.

Compared with RNN and its variants, Convolutional Neural Network (CNN) has no order dependency problem, runs faster, and takes less training time. Therefore, some research work uses CNN to solve aspect-level sentiment analysis task. Xue[11] proposed a model GCAE based on gated convolutional neural networks, which uses CNN for feature extraction, and proposed GTRU gated units (Gated Tanh-ReLU Units) to extract contextual information related to aspect sequences, which achieve a more ideal effect. However, this model only models aspect-specific context information, does not consider the influence of context on specific aspects, that is, it ignores the interaction between context and aspect. Compared with the method based on RNN and its derivative network, the above CNN-based method can achieve parallel processing, spend less time, and can achieve the same or even higher classification accuracy. However, CNN can only obtain the dependency relationship between local words, and cannot obtain the overall structure information of the sentence, which makes the existing
methods still have certain problems.

In response to the above problems, this paper introduces a multi-head attention mechanism based on the GCAE model to capture information closely related to specific aspects in the context sequence, thereby increasing the modeling of specific aspects and making up for the original GCAE model. The insufficiency of contextual information modeling, and then proposed the GCAE-MHA model.

2. The GCAE-MHA model

For convenience, the task definition is given first. Suppose a text sequence \( s = \{w_1, w_2, \ldots, w_i, \ldots, w_n\} \) is given, and its length is \( n \), where \( w_i \) represents the \( i \)-th word in the text sequence. The text sequence may contain one or more aspect sequences \( t = \{w_j, w_{j+1}, \ldots, w_{j+m-1}\} \), where \( m \) is the length of the aspect sequence. The task of aspect-level sentiment analysis is to determine the sentiment polarity of the aspect sequence \( t \) in the text sequence \( s \).

2.1 The improvement ideas

Xue proposed a GCAE model based on a convolutional neural network and proposed a new gating unit GTRU. Neither the convolutional neural network nor the gating unit need to rely on the previous calculation results, so the model can realize parallel computing, which can greatly improve the running speed. The architecture diagram of the GCAE model is shown in Figure 1.

![Figure 1. GCAE model architecture diagram](image)

It can be seen intuitively from Figure 1 that the GCAE model only models the context sequence and ignores the separate modeling of the aspect sequence. Studies have pointed out that in sentiment classification tasks, if the aspect information is not considered, it will cause nearly 40% of errors [12]. Although the GCAE model considers aspect information in the context modeling process, it ignores the interaction between context and aspect information. The research work of Ma shows that in aspect-level sentiment analysis tasks, context and aspect can interact with each other. For reasoning, the context should be aspect-specific, and the aspect should also be context-specific. The two should be modeled separately and the relationship between the two should be learned. This is very helpful for improving the accuracy of aspect-level sentiment analysis.

Therefore, this article introduces the current very popular multi-head attention mechanism on the basis of the GCAE model to increase the modeling of the aspect sequence, so as to learn the interactive relationship between the aspect sequence and the context during the model training process. At the same time, in order to solve the problem that the convolutional neural network can only capture local features,
dilated convolution is introduced to explore the semantic information of the remote context, and the multi-scale context information is extracted through different dilated rates, thereby capturing richer context information.

2.2 Method description
The GCAE-MHA model introduces the multi-head attention mechanism on the basis of the GCAE model, which models the text context and aspect sequence respectively and learns the interaction relationship between them. The model structure is shown in Figure 2.

![Figure 2. GCAE-MHA model structure diagram](image)

The GCAE-MHA model mainly includes four levels, namely the word embedding layer, the feature extraction layer, the interactive learning layer and the sentiment output layer. The function and working mechanism of each layer are described in detail as follows:

(1) Word embedding layer: The role of the word embedding layer is to map each word in the context sequence and the aspect sequence to a continuous value vector representation, and then obtain the corresponding word vector matrix.

(2) Feature extraction layer: The function of the feature extraction layer is to extract important features in the text context sequence matrix and the aspect sequence matrix. For the aspect sequence matrix $E$, a convolution filter $W_d \in \mathbb{R}^{k \times d}$ of size $k \times d$ is used to map $k$ word vector representation into a new feature, and then the filter is translated in the entire aspect matrix to obtain a complete feature sequence. Finally, this feature sequence is passed through relu activation function and the maximum pooling operation obtain the most important features $I$ in the aspect sequence matrix, as shown in equation (1). Among them, $*$ represents the convolution operation, $b$ is the offset.

$$ I = \text{Maxpool} \left( \text{relu} \left( E \ast W_d + b \right) \right) \quad (1) $$

(3) Interactive learning layer: The role of the interactive learning layer is to model the context sequence and the aspect sequence. Extracting the most important information of the two, and to fully learn the relationship between the two during the modeling process.

Context sequence modeling: Because a given text may contain many different aspects, when modeling the context, should be fully consider the important characteristics of different aspects, so as to obtain the context-specific characteristics of the aspect. The GTRU gating unit is still used here. Two convolutional neurons are used to calculate the features of the relu gate and the tanh gate respectively, and then the outputs of the two gates are multiplied bit by bit to obtain aspect-specific context features $d$. The specific calculation process is shown in equations (2),(3) and (4).
\[ a_i = \text{relu}(E'_{i \times k} \ast W^1_c + I + b^1_i) \]  
\[ s_i = \tanh(E'_{i \times k} \ast W^2_c + b^2) \]  
\[ d^i = a_i \times s_i \]

Among them, \( W^1_c \) and \( W^2_c \) represent convolution filters, and \( b^1_i \) and \( b^2_i \) represent offsets. In equation (2), the important aspect feature \( I \) and the context feature are combined and passed to the relu function. Because the value range of the relu function is strictly 0 in the negative direction, and there is no upper limit in the positive direction, so \( a_i \) can be regarded as an aspect feature correlation score with contextual features. Equation (3) is used to extract important contextual features, and then multiply \( a_i \) and \( s_i \) bitwise, so that contextual features that are not related to aspects will be strictly blocked, so as to obtain aspect-specific contextual features, and complete the context sequence modeling.

Aspect sequence modeling: The GCAE-MHA model not only increases the modeling of aspect sequences, but also fully considers the influence of context on aspect sequences in the modeling process. Because each word in the context sequence has a different final representation of the aspect, so the multi-head attention mechanism is used here to extract the content closely related to the aspect sequence in the context and complete the modeling of the aspect sequence. The specific process is as follows:

\[ e^i_j = \frac{\exp(e_i)}{\sum_{j=1}^n \exp(e_j)} \]  
\[ z^i = \sum_{j=1}^n e^i_j x_j \]

Among them, \( e^i \) is the attention weight vector, and the elements in \( e^i \) indicate the degree of correlation between each word in the aspect sequence and each word in the context sequence. Then, the obtained attention weight vector and the original aspect sequence word vector are weighted and summed to obtain the context-specific aspect representation, and the modeling of the aspect sequence is completed. In addition, the multi-head attention mechanism can calculate the representation of words under different linear transformations to obtain more comprehensive attention information, which is of great help to the final emotion classification effect.

Sentiment output layer: The sentiment output layer is used to predict the emotional polarity of a specific aspect. The final representations of the context sequence and the aspect sequence are spliced together to form the final vector representation, and the softmax function is used to output the sentiment polarity, and then the final prediction result is obtained.

3. Experiment

3.1 Dataset

The data set used in the experiment is the most commonly used and authoritative benchmark data set in aspect-level sentiment analysis tasks. They are the user comment data set for Restaurant and Laptop provided by SemEval 2014 Task4[13] and the ACL14 Twitter data set[14]. By analyzing and sorting out the data sets used, the statistics are shown in Table 1.

| dataset    | Positive | Neutral | Negative |
|------------|----------|---------|----------|
| Restaurant-Train | 2164     | 637     | 807      |
| Restaurant-Test  | 728      | 196     | 196      |
| Laptop-Train     | 994      | 464     | 870      |
| Laptop-Test      | 340      | 169     | 128      |
| Twitter-Train    | 1565     | 3127    | 1562     |
| Twitter-Test     | 172      | 346     | 173      |
3.2 Experimental settings

The word vector is initialized with the GloVe word embedding model with a dimension of 300. The learning rate is set to 0.005; the minimum batch size is set to 128; the random inactivation rate dropout is set to 0.5; the number of model iterations epoch is set to 50; and the Adam optimizer is used to update the model parameters. For the context sequence, the convolution kernel size is set to 3, for the aspect sequence, the convolution kernel size is set to 1, and the dilation rate is set to 1, 2, 3. For the multi-head attention mechanism, the number of linear transformations is set to 8.

3.3 Comparison model

In order to verify the effectiveness of the model, we select a number of mainstream aspect-level sentiment analysis methods as the baseline model. It is worth mentioning that most models only use the data set provided by SemEval 2014 Task4 when the article is published, and only use Accuracy as an evaluation indicator. In order to further verify the effectiveness of the model, this article adds the Twitter data set and increases the evaluation indicator F1 value. Therefore, this article reproduces all baseline models, and the hyperparameters used in the experiment are based on their paper. The brief introduction of each baseline model is as follows:

- **LSTM**: Use a unidirectional multi-layer LSTM network to model the context sequence and use it as the basis for sentiment classification, without considering aspect information.
- **TD-LSTM**: Use two independent LSTM networks to model the left and right contexts of the aspect sequence, and then connect their hidden states as the basis for sentiment classification.
- **ATAE-LSTM**: Combine the embedding of the aspect sequence with the vectorized representation of the context sequence, input it into the LSTM network to extract the hidden layer features, and then use the attention mechanism to do a weighted summation of the hidden layer features, which is a typical attention-based mechanism LSTM model.
- **IAN**: Use the LSTM network to model the context and aspect sequences separately, and use the attention mechanism to interactively learn the relationship between the context and the aspect and obtain the final representation, and finally combine the two as the basis for sentiment classification.
- **GCAE**: A method based on gated convolutional neural networks that uses GTRU gated units to capture aspect-specific context features, ignoring the modeling of aspect sequences.
- **IGCN**: Use interactive gated convolutional neural network to model context and aspect sequence without using attention mechanism[15].

3.4 Experimental results and analysis

(1) Comparison and analysis of classification accuracy experimental results

In order to verify the effectiveness of the GCAE-MHA model, we compare the experimental results of each model under the same experimental environment. The comparative experimental results are shown in Table 2.

| Model         | Restaurant Acc | Restaurant F1 | Laptop Acc | Laptop F1 | Twitter Acc | Twitter F1 |
|---------------|----------------|---------------|------------|-----------|-------------|------------|
| LSTM          | 76.35          | 63.05         | 66.07      | 59.38     | 68.11       | 64.73      |
| TD-LSTM       | 77.56          | 64.70         | 68.32      | 61.98     | 69.41       | 68.31      |
| ATAE-LSTM     | 77.88          | 65.90         | 69.21      | 62.47     | 69.88       | 68.53      |
| IAN           | 78.56          | 67.16         | 71.26      | 67.36     | 69.84       | 68.85      |
| GCAE          | 77.48          | 68.11         | 69.46      | 66.52     | 70.35       | 68.64      |
| IGCN          | 79.24          | 68.27         | 71.57      | 67.07     | 70.28       | 68.92      |
| **GCAE-MHA**  | **80.27**      | **69.27**     | **71.04**  | **68.21** | **71.27**   | **69.81**  |

It can be seen from Table 2 that the overall classification effect of the GCAE-MHA model on the three data sets is the best, fully verifying the feasibility and effectiveness of the model. Among them, the LSTM model has the worst performance among all methods. This is because the LSTM model does not consider the aspect information, but takes the text sequence as a whole and directly judges the overall
emotional polarity. The TD-LSTM model considers the position of the aspect sequence, and uses two LSTM networks to model the left and right context of the aspect sequence, which has achieved certain results. The ATAE-LSTM model uses the attention mechanism to focus on important content closely related to the aspect sequence in the context, and stitches the context and aspect embeddings to generate a more reasonable final representation. Therefore, it is superior to the three data sets. The performance of the TD-LSTM model. The IAN model considers the interaction between context and aspect, and achieves better results than the ATAE-LSTM model. It also verifies the importance of modeling the context and aspect sequence and learning the interaction between the two. The GCAE model uses CNN and GTRU gating units to extract aspect-specific context features, and achieves better results than models such as LSTM, TD-LSTM, ATAE-LSTM, etc., but it ignores the interaction between context and aspect, so the overall performance is not as good as the IAN model. The IGCN model uses an interactive gating mechanism to learn aspects and contextual representations, and achieves good results. The GCAE-MHA model has the best effect, because the model not only fully considers the relationship between context and aspect sequence, but also uses the multi-head attention mechanism to capture richer features.

(2) Comparison and analysis of model complexity
The parameter amount and memory usage of the model can directly reflect the complexity of the model, so this article analyzes the complexity of different models through a set of comparative experiments. In the same experimental environment, the data set uses Restaurant, the minimum batch size is set to 64, and the parameters and memory usage of each model are counted. The results are shown in Table 3.

| Model       | Params (×10^6) | Memory (MB) |
|-------------|----------------|-------------|
| LSTM        | 0.73           | 10.25       |
| TD-LSTM     | 1.46           | 12.23       |
| ATAE-LSTM   | 2.56           | 16.36       |
| IAN         | 2.19           | 15.04       |
| GCAE        | **0.68**       | **9.21**    |
| IGCN        | 0.93           | 9.81        |
| BGCN        | 1.17           | 10.31       |

It can be seen from Table 3 that the parameter amount and memory usage of the GCAE-MHA model are relatively small, which fully verifies the simplicity and lightweight of the model. On the whole, the model parameters and memory occupancy of the LSTM network are the most. This is because the LSTM network itself is more complex, has a large number of parameters, and requires a lot of memory to save all hidden states. The GCAE model based on convolutional neural network is relatively simple, so its parameter amount and memory usage are less. On this basis, the IGCN model adds the modeling of aspect sequences, so its parameter amount has increased, and the memory usage is correspondingly increased. The GCAE-MHA model also adds the modeling of the aspect sequence, but it is realized by the multi-head attention mechanism, which is more complicated than the gating unit, so its parameter memory occupies more than the IGCN model, but GCAE- should be considered. The MHA model has excellent performance in improving the accuracy of sentiment classification, and the parameter amount and memory usage of the model are much lower than the model based on LSTM.

Based on the above experimental results, the GCAE-MHA model can improve the accuracy of aspect-level sentiment analysis while ensuring the simplicity of the model, fully verifying the feasibility of the design idea of this model.

4 .Conclusion
In the face of massive text data, aspect-level sentiment analysis technology can more effectively help people obtain the huge value contained therein, which has important practical significance and research
value. Aiming at the problem of GCAE model ignoring aspect sequence modeling, the multi-head attention mechanism is used to model aspect sequence on the basis of GACE model, and the interaction relationship between the two is learned, and then the GCAE-MHA model is proposed. Finally, through experiments, a comprehensive comparison and analysis of the BGCN model in terms of emotion classification accuracy, F1 value, and model complexity, fully verified the feasibility and effectiveness of its design ideas. In future work, we can consider applying some excellent pre-training models to our model, such as the BERT model[16], to further improve its effectiveness.

References
[1] Fatemeh H, Mohammad K S.(2019) A survey on classification techniques for opinion mining and sentiment analysis. Artificial Intelligence Reviewe, 52(3):1495-1545.
[2] Thet T T, Na J C, Khoo C S. (2010)Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science, 36(6): 823–848.
[3] Tang D, Qin B, Feng X. (2015)Effective LSTMs for Target-Dependent Sentiment Classification. Computer Science, 2015: 3298-3307.
[4] Mnih V, Heess N, Graves A. (2014) Recurrent models of visual attention. In: The 28th Conference on Neural Information Processing Systems. BALI. 2204-2212.
[5] Cheng Y, Shen S Q, He Z J. (2016) Agreement-based joint training for bidirectional attention-based neural machine translation. In: The 25th International Joint Conference on Artificial Intelligence. Manhattan. 2761-2767.
[6] Mohammad E, Shahla N, Moloud A. (2021) ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis. Future Generation Computer Systems, 115:279-294.
[7] Wang Y, Huang M, Zhu X. (2016) Attention-based LSTM for Aspect-level Sentiment Classification. In: the 2016 conference on empirical methods in natural language processing. Austin. 606-615.
[8] Ma D H, Li S J, Zhang X D. (2017) Interactive Attention Networks for Aspect-Level Sentiment Classification. In: IJCAI. Melbourne. 4068-4074.
[9] Huang Q, Ou Y, Carley K M. (2018) Aspect Level Sentiment Classification with Attention-over-Attention Neural Networks. In: the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer. Washington DC. 197-206.
[10] Yiming C, Zhipeng C, Si W. (2017) Attention-over-Attention Neural Networks for Reading Comprehension. In: ACL. Vancouver. 593-602.
[11] Xue W, Li T. (2018) Aspect based sentiment analysis with gated convolutional networks. In: the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne. 2514-2523.
[12] Jiang L, Yu M, Zhou M. (2011) Target-dependent twitter sentiment classification. In: the 49th Annual Meeting of the Association for Computational Linguistics. Portland. 151-160.
[13] Pontiki M, Galanis D, Pavlopoulos J. (2014) SemEval-2014 Task 4: aspect based sentiment analysis. In: the 8th International Workshop on Semantic Evaluation. 27-35.
[14] Dong L, Wei F, Tan C. (2014) Adaptive recursive neural network for target-dependent twitter sentiment classification. In: the 52nd Annual Meeting of the Association for Computational Linguistics. ACL Press, 49-54.
[15] Avinash K, Vishnu T, Narapareddy. (2020) Aspect-Based Sentiment Classification Using Interactive Gated Convolutional Network. IEEE Access. 8: 22445-22453.
[16] Devlin J, Mingwei Chang, Kenton Lee. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL. Florence. 4171-4186.