A Optimized BERT for Multimodal Sentiment Analysis

JUN WU, TIANLIANG ZHU, JIAHUI ZHU, TIANYI LI, and CHUNZHI WANG, School of Computer Science, Hubei University of Technology, China

Sentiment analysis of one modality (e.g., text or image) has been broadly studied. However, not much attention has been paid to the sentiment analysis of multi-modal data. As the research on and applications of multi-modal data analysis are becoming more and more broad, it is necessary to optimize BERT internal structure. This article proposes a hierarchical multi-head self-attention and gate channel BERT, which is an optimized BERT model. The model is composed of three modules: the hierarchical multi-head self-attention module realizes the hierarchical extraction process of features; the gate channel module replaces BERT's original Feed Forward layer to realize information filtering; and the tensor fusion model based on a self-attention mechanism is utilized to implement the fusion process of different modal features. Experiments show that our method achieves promising results and improves accuracy by 5–6% when compared with traditional models on the CMU-MOSI dataset.

CCS Concepts: • **Information systems** → *Multimedia information systems;*

Additional Key Words and Phrases: HG-BERT, multi-head self-attention mechanism, multimodal sentiment analysis, gate channel, tensor fusion network

ACM Reference format:
Jun Wu, Tianliang Zhu, Jiahui Zhu, Tianyi Li, and Chunzhi Wang. 2023. A Optimized BERT for Multimodal Sentiment Analysis. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 2s, Article 91 (February 2023), 12 pages.
https://doi.org/10.1145/3566126

1 INTRODUCTION

Since the development of social media, there are billions of multimedia data, including text, audio, video, and so on. The complexity of emotional information is increasing, and the valuable information is also increasing. Therefore, mining emotional characteristic information can not only uncover people's position or attitude toward some hot topics or important events [10] but also carry out customized recommendation services. So there is a major challenge regarding how to process and analyze these multi-modal data.

Jun Wu and Tianliang Zhu contributed equally to this research.

This work is supported by the National Natural Science Foundation of China (Grant No. 61602161, 61772180), Hubei Province Science and Technology Support Project (Grant No: 2020BAB012), The Fundamental Research Funds for the Research Fund of Key Lab of Traffic and Internet of Things (WUT:2015-015-A03).

Authors' address: J. Wu, T. Zhu, J. Zhu, T. Li, and C. Wang, School of Computer Science, Hubei University of Technology, Wuhan, Hubei, China 430068; emails: wujun@whut.edu.cn, 913829070@qq.com, 314328625@qq.com, 469216198@qq.com, lavazza@foxmail.com.

1551-6857/2023/02-ART91 $15.00
https://doi.org/10.1145/3566126

In recent years, with the development of deep learning, more and more researchers have shifted their attention from traditional machine learning to using deep learning to process natural language processing. Pu et al. [19] used Support Vector Machine to deal with emotion classifications based on the document level. Some researchers use reinforcement learning to deal with affective categorization. Chen et al. [4] uses it to calculate the emotional value of words and obtains the emotion of sentences by accumulating the emotional value of words. Some researchers also use reinforcement learning to control information input by use of a gating mechanism [2]. These methods provide a new way of thinking, which uses a reinforcement learning algorithm for affective categorization, but it is difficult to take advantage of reinforcement learning in decision making.

For the study of text mode and other modes (video or audio, etc.), a **Long Short Term Memory (LSTM)** [9] model is used more often because of its simple structure and different gating mechanisms to control the output of features, thus alleviating the problem of gradient disappearance caused by **Recurrent Neural Network (RNN)**. Therefore, LSTM is also widely used in data modes with sequence characteristics. Zadeh [29] uses LSTM to process multi-modal data (audio, text, and vision) and uses tensor fusion to realize the fusion of different data features. Liu [16] degraded the three-dimensional feature matrix by optimizing the fused features, so as to achieve high efficiency in the operation process. Similarly, the **Convoluted Neural Networks (CNN)** network is widely used in the image field because of the ability to extract local features efficiently using different convolutional kernels. Liu [14] proposed to integrate adversarial learning into the invariant image text learning mode. For the specific task of image sentence matching, CNN-RNN was used to embed the network. Similarly, in feature fusion, CNN can be used to control the number of convolution to unify the dimensions of features of different modal data [28]. Since it lacks the ability to process sequence features, it is often used as the auxiliary structure of a network model to make up for the deficiency of network model.

With the emergence of Transformer [24], **Bidirectional Encoder Representation from Transformers (BERT)** [6] and other pre-training models, its powerful feature extraction ability has achieved SOTA results in 11 NLP tasks. Transformer solves the inefficiency of LSTM, which processes serial data by using location coding to realize parallel data processing. At the same time, the key part of the data features can be found by multi-head self-attention mechanism. Using a powerful pre-training process, each word is given a wealth of information. Various models based on Transformer, such as BERT and A Robustly Optimized BERT Pretraining Approach [15], have achieved good results by using bidirectional pre-training process and increase more number of parameters. Regarding BERT and other pre-training models as text features extraction, using auxiliary structures such as LSTM to extract features of other modes or proposing new feature fusion methods has become a hot topic in the task of studying multimodal emotion classification.

However, most of the later research methods based on multimodal sentiment analysis are based on fine-tuning the BERT model, and there are few models that directly optimize BERT. As shown in Figure 1, we propose the **hierarchical multi-head self-attention and gate channel BERT (HG-BERT)** model. The following innovations are proposed:

- `Hierarchical multi-attention mechanism`: This is used to realize hierarchical extraction of data features.
- `Gate channel`: This is utilized to replace the Feed Forward layer in the BERT model to realize noise filtering.
- `HG-Bert model`: This is the optimized BERT model with hierarchical multi-attention mechanism and gate channel.
- `Feature Fusion`: Information interaction between models is realized through a tensor fusion model based on self-attention.

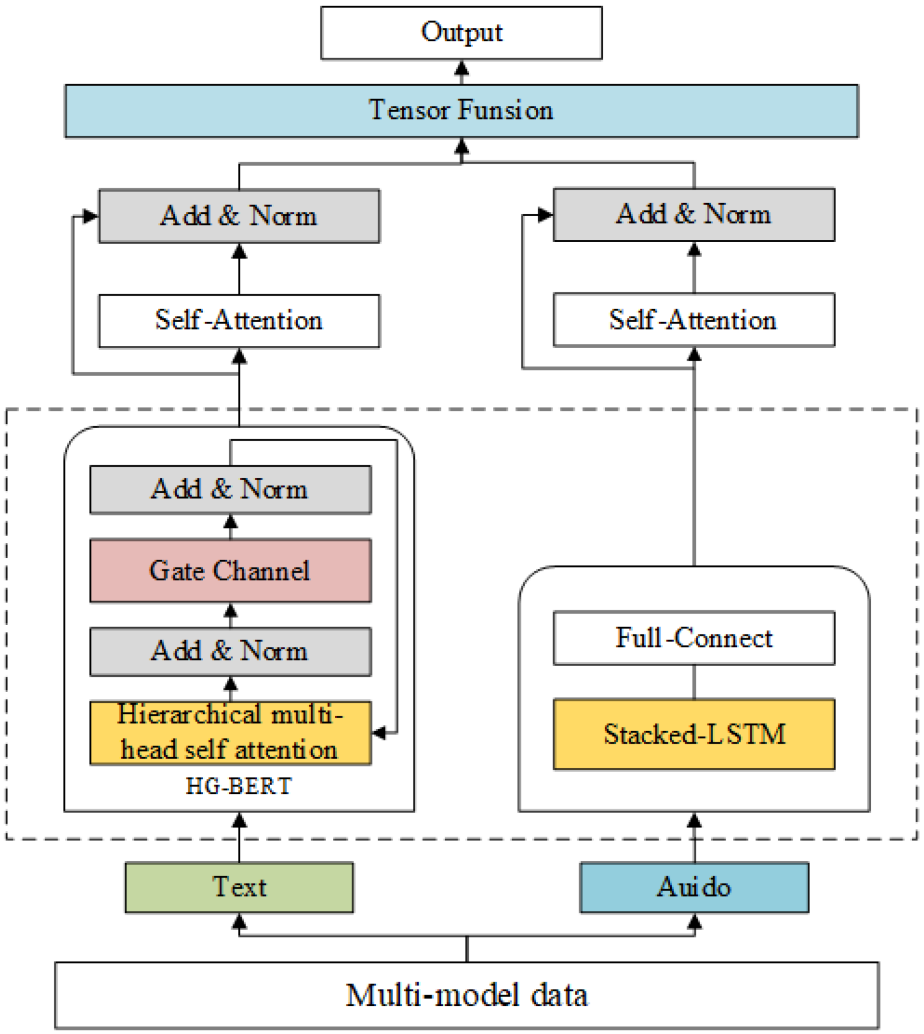

Fig. 1. Modal fusion process of text and audio. It consists of an optimized BERT model and a custom multi-modal feature fusion method.

2 RELATED WORK

2.1 Multimodal Sentiment Analysis Model

For multi-modal feature extraction models from traditional LSTM to Transformer and BERT, the classification effect is better, but the number of parameters and training time are also increasing. Researchers need to weigh the number of parameters required for training against the classification results achieved. At the same time, the network model should be selected according to the characteristics of data. For example, using CNN for image-based data will work well. Usama [23] proposed a new model based on RNN with a CNN-based attention mechanism. CNN learns the high-level features of sentence from input representation, and RNN was used to deal with the features processed by CNN. Better results are achieved by combining LSTM and Transformer's multi-head self-attention mechanism [11]. Liu [13] proposes a multi-classification sentiment analysis model that concatenates the sentiment features extracted by CNN and LSTM to express

the sentiment features of the text. Sun [20] propose a model based on LSTM and the attention mechanism to predict the sentiment of each target. Due to the simplicity of LSTM structure, it is difficult to directly and deeply extract features.

Transformer improves RNN's most criticized shortcoming of slow training and uses a self-attention mechanism to achieve fast parallelism. And Transformer can be increased to a very deep depth, fully exploiting the characteristics of the RNN model and improving the accuracy of the model. The follow-up models Roberta, Decoding-enhanced BERT with disentangled attention (Deberta) [8], and so on, achieved the SOTA effect by increasing the parameter amount of the pre-training model and optimizing the structure of Embedding. This experiment is also trained according to the general structure of Transformer. But after understanding the structure of Transformer, we first put forward our own ideas on the structure of Feed Forward and multi-head self-made attention layer in its structure. In the Feed Forward layer, the Transformer uses a fully connected layer to implement dimensional changes to the features. We improve the Forward Feed layer and s similar GRU structure to filter features rather than a simple full connet layer.

On the fusion of multimodal features, the CM_BERT model uses masked multimodal attention to dynamically adjust the weight of words by combining information from text and audio modalities. However, in the fusion process, only the audio feature is used as a weighting factor to measure the information of the text feature. Audio feature information is ignored to some extent. Xu [27] propose a Cross-Modal Hybrid Feature Fusion framework that can directly learn the image–text similarity by fusing multimodal features with inter- and intra-modality relations incorporated. In the process of multimodal feature fusion, this article draws on part of the masked multimodal attention process and realizes the preservation of different modal feature information through the tensor fusion model.

2.2 Improvement of the Optimization Model Based on BERT

At the end of model training, a set of weight values with good results is obtained, and this set of weight values is shared with others, which is called the pre-training model. In recent years, pretraining models have been widely used in multimodal emotion classification tasks. A strong pre-training model has a lot of room for improvement. In previous studies, the LSTM used previous semantic information to infer current information. ELMO [18] solved the polysemy problem by using bi-directional LSTM to construct text information. BERT used MLM and NSP tasks for the pre-training process. The mask model is used to achieve the purpose of deep bidirectional joint. An NSP method is used to capture the dichotomous task between sentences. RoBERTa [1] uses dynamic mask to train the model and uses BPT (based on byte encoding) to process text information. Deberta [8] proposed a disentangled attention mechanism, which represented a word by using two vectors that encode its content and position. In addition, a new adversarial training method is proposed for fine-tuning to improve models' generalization. StructBert [25] explicitly model language structures by forcing the model to reconstruct the right order of words and sentences for correct prediction by incorporating language structures into pre-training.

3 HIERARCHICAL MULTI-HEAD SELF-ATTENTION AND GATE CHANNEL BERT MODEL

Due to BERT's strong feature extraction ability, most studies are fine-tuned based on the BERT model. BERT is used to extract features, and other simple models are used to process the features extracted by BERT. In the process of studying BERT, it is found that some network structures in BERT can be optimized, so the HG-BERT model is proposed. Compared with the original BERT model, the improvement points of HG-BERT are shown in Figure 2, which can be divided into three aspects: First, the hierarchical multi-head self-attention mechanism process is used, and sec-

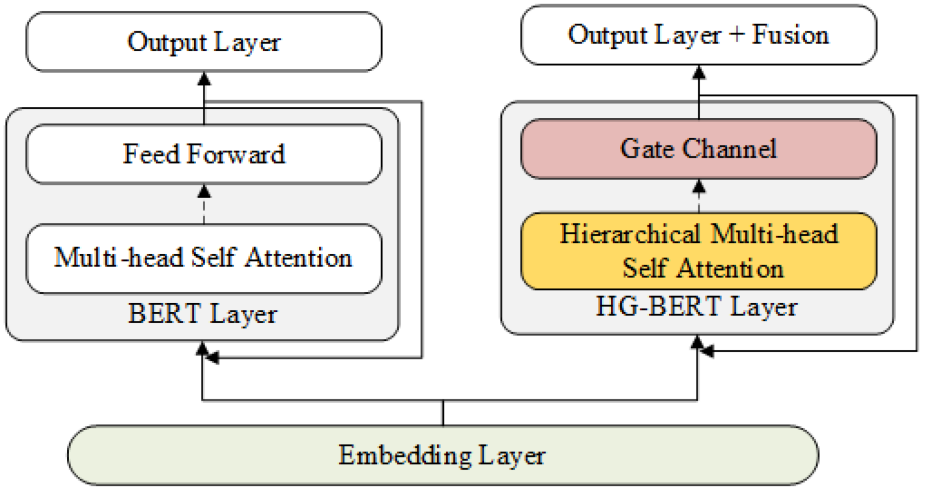

Fig. 2. BERT (left) and the HG-BERT (right) model. HG-BERT changes two contents on the basis of original BERT, one was multi-head Self-Attention and the other was Feed Forward. In the fusion network, it is improved on the basis of tensor fusion network.

ond, the feed forward layer in the BERT model is replaced by a gate channel. Finally, according to the fusion process of multi-modal emotion data, a fusion method based on self-attention is proposed.

3.1 Hierarchical Multi-head Self-attention

BERT uses a fixed number of heads to process embedded data, and the extraction process of data features is the same. However, similarly to the working principle of CNN image processing, different features can be obtained when processing the same data with different head numbers. Dealing with different numbers of heads means that the dimension of the data varies during self-attention. The extraction ability of data features is different for different layers. Therefore, a variable number of self-attentional headers is proposed, as shown in Figure 3. Different head sizes are used at different levels. At the beginning of the Bert layer, the number of heads is divided for extensive processing of local features of data. In the last few layers, because the data characteristics have already been filtered by the previous layers, only a small number of headers need to be set later. In the experiment, the head number distribution of hierarchical multi-head self-attention mechanism is as follows: (1) the first 1–3 layers are set as head_num = 16; (2) the first 4–6 layers are set as head_num = 12; (3) the first 7–9 layers are set as head_num = 8, and the last 10, 11, and 12 layers are set as head_num = 4. After the BERT embedding layer, its feature dimensions are $X \in R^{16\times48}$(layers 1–3), X $\in R^{12\times64}$(layers 4–6), X $\in R^{8\times96}$(layers 7–9), and X $\in R^{4\times192}$(layers 10–12).

3.2 Gated Information Channel

Sentiment analysis of one modality (e.g., text or image) has been broadly studied. However, not much attention has been paid to the sentiment analysis of multimodal data. As the research on and applications of multimodal data analysis have become more broad, it is necessary to work on sentiment by combining the visual content with text descriptions. In 2020, Feiran Huang [10] proposed a novel method, Attention-based Modality-Gated Networks, to exploit the correlation between the modalities of images and texts and extract the discriminative features for multimodal sentiment analysis. Similarly to the network structure of GRU [5], this article designs a similar

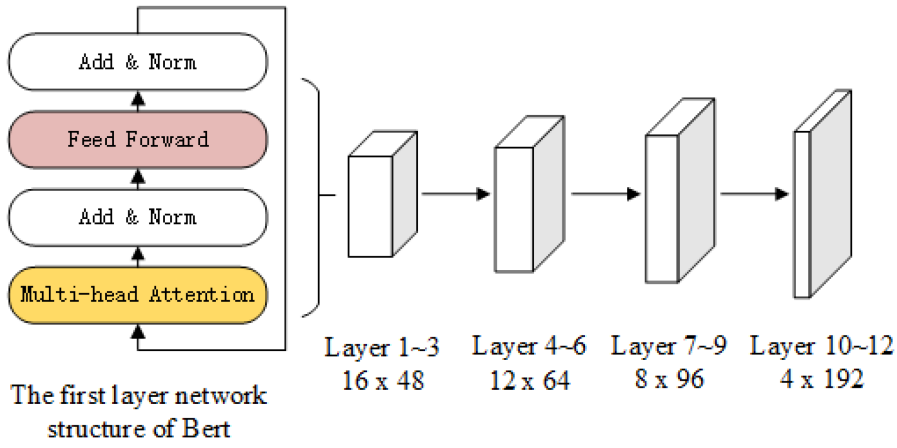

Fig. 3. Hierarchical multi-head self-attention mechanism structure. Use different numbers of heads in different layers to process features.

gated information mechanism. As shown in Figure 4, the gate information channel consists of two parts: one is the memory gate and the other is the update gate. Memory gates are used to hold valuable information, while additional channels are added to remember new information. We use another update gate to implement updates to features:

$$G\_M = \sigma(X \otimes M), \tag{1}$$

$$G\_U = \sigma(X \otimes U), \tag{2}$$

X($X \in R^{b \times len \times hn}$) is the feature after hierarchical multi-head self-attention mechanism; M and U are the parameter matrix $M, U \in R^{hn \times hn}$; b represents batch size; len represents the fixed length of text; and hn represents feature dimension. Memory gates are used to store information and add some new content to features as follows:

$$X\_M = \tanh(X \otimes G\_M + X \otimes M). \tag{3}$$

Next, we use the update gate to update the original data as follows:

$$X_{new} = X\_M \otimes G\_U + X \otimes (1 - G\_U). \tag{4}$$

Finally, after the data feature X is obtained, a residual network and BERT LayerNorm are used for the final data update.

3.3 Tensor Fusion Method Based on Self-Attention

The processing process can be divided into two stages: The first stage is to use the self-attention mechanism to deal with data features, and the second stage is the data fusion process in Figure 5. The feature extraction for text data is as follows: First, the improved HG-BERT is used to process text information. A self-attention mechanism is used to find important parts of data, and a residual network and BERT layer regularization are used to standardize features. For feature processing of audio data, we first use stacked double-layer LSTM to process audio features. Which has better effect. After using the self-attention mechanism and layer regularization, the last feature vector is used as the representation of audio data in the output obtained. At the tensor fusion step, the fused features pass through the full connection layer, which is convenient to combine with the original features again.

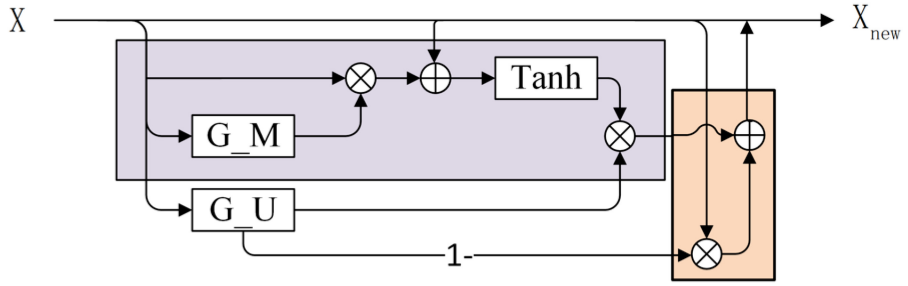

Fig. 4. Structure of gate channel. Through different gate structures G_M and G_U, feature X can be filtered. The G_M implementation stores useful information about the feature. G_U implements a feature update. Finally, the residual structure is used to obtain the final output.

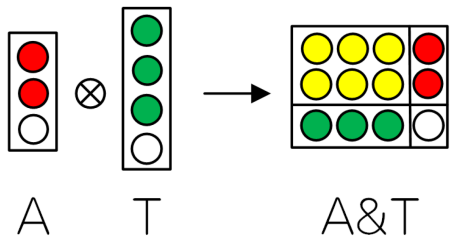

Fig. 5. Modal fusion process of text and audio. Through the text feature of the BERT network, tensor fusion with audio feature is carried out.

In the process of multi-modal data feature processing, tensor fusion model it is used to store information about other modes by adding an extra dimension to each single modal feature. The fusion features not only have the single modal features before the fusion but also have cross-modal information between different modes. Compared with single mode, it has more information.

4 ANALYSIS OF EXPERIMENTAL RESULTS

4.1 Experimental Environment and Dataset

To test the HG-BERT model that has been designed in Section 3, it is necessary to determine whether the optimized model performed better than the baselines. First, we introduce the operating environment of the experiment, as shown in as Table 1.

Next, we choose CMU-MOSI [30] as the dataset of our experiment that could provide both raw and processed data characteristics. Here the CMU-MOSI open dataset processed in the paper [28] CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis.

4.2 Baseline

We compared the performance of HG-BERT with the following experiment, and the model to be compared is as follows:

Tensor Fusion Network (TFN) [29]: This model uses cross-products of the features of multiple modes, adding an extra space to each feature mode and preserving the features of multiple different dimensions.

Low-Rank Multimodal Fusion (LMF) [16]: This model uses low-rank weight tensors to efficiently decompose the networks that perform tensor cross-product.

Table 1. Experimental Platform and Parameter Setting

Experimental Environment	Configuration
Operating system	Window10
Processor	AMD Ryzen 7 4800H with Radeon Graphics 2.90 GHz
Memory	8.00 GB
Torch Version	torch1.10.0+cu102
Programming Language	Python3.6
Deep learning Framework	Pytorch

Gate Multimodal Embedding-LSTM (GME-LSTM) [3]: This model uses reinforcement learning to control the information of the gating mechanism and alleviates the influence of noise in feature fusion.

Multimodal Factorization Model (MFM) [22]: This model optimizes for a joint generative discriminative objective across multimodal data and labels by factorizing representations into two sets of independent factors: multimodal discriminative and modality-specific generative factors.

Recurrent Multistage Fusion Network (RMFN) [12]: This model breaks down the fusion problem into multiple phases that focu on a subset of multiple patterns.

Multimodal Cyclic Translation Network (MCTN) [7]: This model forms a method of learning joint representations basing on the important point that translation from a source to a target modality.

Deberta [8]: This model proposed a disentangled attention mechanism that represented a word by using two vectors that encode its content and position. In addition, a new virtual adversarial training method is proposed for fine-tuning to improve models' generalization.

Multimodal Transformer (MulT) [21]: This model uses directional pairwise cross-modal attention that attends to interactions between multimodal sequences across distinct time steps and latently adapt streams from one modality to another.

CM-BERT+: This model uses the original CM-BERT [28] model to extract text features and a tensor fusion method based on self-attention.

BERT-TA+: This model uses the original BERT [6] model to extract text features and a tensor fusion method based on self-attention.

4.3 Results Analysis

The experiment will evaluate the HG-BERT model on the CMU-MOSI dataset, and the experimental results are shown in Table 2. The experimental evaluation criteria for model performance were Accuracy and F1_Score. F1_Score is the harmonic average of accuracy and recall, with a maximum of 1 and a min- imum of 0. In the experiment, the f1_score function in the SKLearn library is used to calculate F1_Score, and a weighting method is adopted.

First, in the processing of multi-modality data based on LSTM, TFN uses a tensor cross-product to fuse feature vectors of different modes. Since it is in the form of a cross-product of feature vectors, the amount of feature data resulting from fusion is N^3. LMF degrades the tensor matrix on the basis of the TFN model. Both models use the traditional LSTM model, and the new added step is feature fusion. As a result, the experimental effect is not good, and its accuracy is 77.1% and 76.4%, respectively. Based on LSTM, GME-LSTM introduces reinforcement learning to update the gated information. Since the decision classification problem, at which reinforcement learning excels, is not introduced, its accuracy is 76.5%. MFM optimizes the joint generation identification model

Table 2. The Experimental Results of Different Models on the CMU-MOSI Dataset

Model	Modality	ACC/%	F1
TFN [29]	T+A+V	77.1	
LMF [16]	T+A+V	76.4	75.7
GME-LSTM [3]	T+A+V	76.5	
MFM [22]	T+A+V	78.1	78.1
RMFN [12]	T+A+V	78.4	78.0
MCTN [7]	T+A+V	79.3	79.1
Deberta+ [8]	T	81.9	81.9
MulT [21]	T+A+V	83.0	82.8
CM-BERT+	T+A	82.65	82.64
BERT-TA+	T+A	83.38	83.45
HG-BERT (ours)	**T+A**	**83.82**	**83.91**

"+" represents the code we duplicated.

in cross-modal data. Generative factors are shared across channels and contain joint multimodal characteristics, and discriminant factors contain information about generated data. RMFN divides the fused features into different stages, focusing only on a small number of features in each stage. MCTN is a learning method of joint feature presentation between different modes. These models all bring forward the innovation of feature fusion, but they do not improve the degree of feature extraction. Deberta achieves an accuracy of 81.8% by using the disentangled attention mechanism, but its accuracy will be better with different parameter configurations. For example, when only changing the learning rate to 2e-5, the effect will be 0.1% better than the original. The MulT model uses a bidirectional cross-modal attention mechanism that enables learning at each time step. By improving the training model, the accuracy is improved. When using BERT's model, the results of CM-BERT+, the mask attention mechanism model, and BERT-TA+, the two-mode tensor fusion model based on self-attention, reached 82.65% and 83.38%, respectively, which was 3–5% higher than that of LSTM. This shows that the BERT model is powerful for feature extraction. Finally, by optimizing the BERT model, the accuracy of the HG-Bert model proposed is improved by 0.44% compared with the BERT model, which provides an idea for further research on the optimization BERT model.

4.4 Effectiveness of Multi-modal Explanations

For further comparison and analysis among the BERT models and the HG-BERT model, it is necessary to design the group experiments, and their experimental results are shown in Table 3.

Next, a BRET model was used, and its accuracy was 83.67% (Group A) and 83.38% (Group B3) when using single mode (text) and double mode (text and audio), respectively. The single-mode performance of the experiment is about 0.3% higher than that of the double mode, indicating that the audio data become noise during the fusion, which affects the classification effect of the BERT model. There may be two reasons for this: First, more noise is mixed in audio data extraction and word-based alignment, which affects the accuracy of audio data. Second, the processing of audio data using LSTM is not sufficient. Next, we demonstrate the effectiveness of the third innovation proposed in this article. For the comparison of fusion methods, the mask attention fusion method (Group B4, CM-BERT) was used in the study. Under the condition of ensuring the same parameter configuration as used in the article, the experimental effect of tensor fusion method (Group B3) based on self-attention was 1% better than that of B4. At the end of the fusion method of mask

Table 3. Ablation Research of the HG-BERT

Group	Model	Modality	ACC/%	F1/%
A	BERT	T	83.67	83.70
B1	A + hierarchical multi-head attention (HM)	T	82.07	82.08
B2	A + gate channel (GC)	T	82.22	82.17
B3	A + tensor fusion based on self-attention (TFS)	T + A	83.38	83.45
B4	A + fusion (CM-BERT)	T + A	82.36	82.32
C1	A + GC + TFS	T + A	79.1	79.0
C2	A + HM + TFS	T + A	82.8	82.77
C3	A + HM + GC	T + A	82.22	82.27
D	A + HM + GC + TFS	T + A	83.82	83.91

Table 4. The Result of Head Distribution in Hierarchical Multi-head Self-attention Mechanism

Group	Model	Learning Rate	Head Distribution	Modality	ACC/%	F1/%
E1	A(BERT)	1e-5	12-12-12-12	T	83.67	83.68
E2	A+LM	1e-5	16-12-8-4	T	82.07	82.08
E3	A+LM	1e-5	16-12-12-8	T	83.53	83.56
F1	A(BERT)	2e-5	12-12-12-12	T	82.22	82.21
F2	A+LM	2e-5	16-12-8-4	T	82.22	82.19
F3	A+LM	2e-5	16-12-12-8	T	82.65	82.64

attention, the attention coefficient obtained from audio and text is used to score with the original text features, which ignores the audio data features. In this article, the tensor fusion method is used to preserve not only the information of original modes but also the information of interaction between modes.

Finally, the HG-BERT module is tested experimentally, and the experimental results are relatively the same (lower than the highest accuracy of about 1%). It had a poor effect only in the C1 experiment. The reason may be that the design of the gate channel module is not good enough. The experiment might be better if it were modified. When carrying out the influence of head distribution on the BERT model in the multi-modality self-attention mechanism, the experiment tests the experimental effect of head distribution, which is the first proposed innovation in this article, and the experimental results are shown in Table 4.

Experiments are conducted to compare the accuracy of the BERT model and the hierarchical multi-head self-attention mechanism under different learning rates based on BERT. In groups E and F, the difference between the test results and the original BERT result was 1.6%. Group F1 and F3 on Table 4 which used the same head distribution results are slightly better than BERT in Group F1, using the same head distribution results are slightly better than BERT. Experimental results show that head distribution in BERT is one of the performance indicators that can affect BERT, and the performance of BERT model can be improved by adjusting head distribution.

5 CONCLUSION

This article proposed an optimized model based on BERT. First, a hierarchical multi-head self-attention mechanism is used to extract feature by using a progressive number of head, taking advantage of the difference of feature extraction capability of different BERT network layers. Second, the gate channel is bought into the BERT model for filter information. Third, the self-attention mechanism is used for multi-mode fusion. The HG-BERT model optimized as the above steps has improved, and its experiment result on CMU-MOSI [30] got better than the traditional models.

There are similar datasets, CMU-MOSEI [2], YouTube [17], ICT-MMMO [26], and so on. All of them are useful to go further with experiment and research. However, there are still some problems in our experiment. In the process of using hierarchical multi-head self- attention, a manually specified head number distribution is used. In future research, network training can be used to obtain parameters to specify its distribution. Second, in the design process of the gate mechanism, due to the limited research level, it is difficult to give a theoretical explanation, which can be further discussed in the future research studies.

6 ETHICS APPROVAL

Our studies present no ethical issues.

7 CONFLICT OF INTEREST

All authors declare that we have no conflict of interest.

REFERENCES

[1] Mehdi Arjmand, Mohammad Javad Dousti, and Hadi Moradi. 2021. TEASEL: A transformer-based speech-prefixed language model. https://arxiv.org/abs/2109.05522.

[2] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2236–2246. https://doi.org/10.18653/v1/P18-1208

[3] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI'17)*. Association for Computing Machinery, New York, NY, 163–171. https://doi.org/10.1145/3136755.3136801

[4] Ruiqi Chen, Yanquan Zhou, Liujie Zhang, and Xiuyu Duan. 2019. Word-level sentiment analysis with reinforcement learning. *IOP Conf. Ser.: Mater. Sci. Eng.* 490, 6 (2019), 062063. https://doi.org/10.1088/1757-899x/490/6/062063

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078. Retrieved from https://arxiv.org/abs/1406.1078.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:cs.CL/1810.04805. Retrieved from https://arxiv.org/abs/1810.04805.

[7] Pham Hai, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6892–6899. https://doi.org/10.1609/aaai.v33i01.33016892

[8] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. arxiv:2006.03654. Retrieved from https://arxiv.org/abs/2006.03654.

[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[10] Feiran Huang, Kaimin Wei, Jian Weng, and Zhoujun Li. 2020. Attention-based modality-gated networks for image-text sentiment analysis. *ACM Trans. Multimedia Comput. Commun. Appl.* 16, 3, Article 79 (July 2020), 19 pages. https://doi.org/10.1145/3388861

[11] Xue-Liang Leng, Xiao-Ai Miao, and Tao Liu. 2021. Using recurrent neural network structure with enhanced multi-head self-attention for sentiment analysis. *Multimedia Tools Appl.* 80, 8 (2021), 12581–12600. https://doi.org/10.1007/s11042-020-10336-3

[12] Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. arXiv:cs.LG/1808.03920. Retrieved from https://arxiv.org/abs/1808.03920.

[13] Lei Liu, Hao Chen, and Yinghong Sun. 2021. A multi-classification sentiment analysis model of chinese short text based on gated linear units and attention mechanism. *ACM Trans. As. Low-Resour. Lang. Inf. Process.* 20, 6, Article 109 (Sep. 2021), 13 pages. https://doi.org/10.1145/3464425

[14] Ruoyu Liu, Yao Zhao, Shikui Wei, Liang Zheng, and Yi Yang. 2019. Modality-invariant image-text embedding for image-sentence matching. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1, Article 27 (February 2019), 19 pages. https://doi.org/10.1145/3300939

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:cs.CL/1907.11692. Retrieved from https://arxiv.org/abs/1907.11692.

[16] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank multimodal fusion with modality-specific factors. arXiv:cs.AI/1806.00064. Retrieved from https://arxiv.org/abs/1806.00064.

[17] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI'11)*. Association for Computing Machinery, New York, NY, 169–176. https://doi.org/10.1145/2070481.2070509

[18] Matthew Peters, M. Neumann, M. Iyyer, M. Gardner, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. https://doi.org/10.18653/v1/N18-1202

[19] Xiaojia Pu, Gangshan Wu, and Chunfeng Yuan. 2019. Exploring overall opinions for document level sentiment classification with structural SVM. *Multimedia Syst.* 25, 1 (2019), 21–33. https://doi.org/10.1145/161468.161469

[20] Chengai Sun, Liangyu Lv, Gang Tian, and Tailu Liu. 2020. Deep interactive memory network for aspect-level sentiment analysis. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 20, 1, Article 3 (December 2020), 12 pages. https://doi.org/10.1145/3402886

[21] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Conference of the Association for Computational Linguistics*, Vol. 2019. NIH Public Access, 6558.

[22] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning Factorized Multimodal Representations. arXiv:cs.LG/1806.06176. Retrieved from https://arxiv.org/abs/1806.06176.

[23] Mohd Usama, Belal Ahmad, Enmin Song, M. Shamim Hossain, Mubarak Alrashoud, and Ghulam Muhammad. 2020. Attention-based sentiment analysis using convolutional and recurrent neural network. *Fut. Gener. Comput. Syst.* 113 (2020), 571–578. https://doi.org/10.1016/j.future.2020.07.022

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

[25] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. StructBERT: Incorporating language structures into pre-training for deep language understanding. Arxiv.1908.04577. Retrieved from https://arxiv.org/abs/1908.04577.

[26] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. YouTube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intell. Syst.* 28, 3 (2013), 46–53. https://doi.org/10.1109/MIS.2013.34

[27] Xing Xu, Yifan Wang, Yixuan He, Yang Yang, Alan Hanjalic, and Heng Tao Shen. 2021. Cross-modal hybrid feature fusion for image-sentence matching. *ACM Trans. Multimedia Comput. Commun. Appl.* 17, 4, Article 127 (November 2021), 23 pages. https://doi.org/10.1145/3458281

[28] Kaicheng Yang, Hua Xu, and Kai Gao. 2020. *CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis*. Association for Computing Machinery, New York, NY, 521–528. https://doi.org/10.1145/3394171.3413690

[29] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. arXiv:cs.CL/1707.07250. Retrieved from https://arxiv.org/abs/1707.07250.

[30] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv:cs.CL/1606.06259. Retrieved from https://arxiv.org/abs/1606.06259.

Received 3 March 2022; revised 9 June 2022; accepted 16 August 2022