applied sciences

MDPI

Article

Multi-Modal Sentiment Analysis Based on Interactive Attention Mechanism

Jun Wu, Tianliang Zhu, Xinli Zheng * and Chunzhi Wang

School of Computer Science, Hubei University of Technology, Wuhan 430068, China
* Correspondence: 102101053@mail.hbut.edu.cn; Tel.: +86-027-59750444

Abstract: In recent years, multi-modal sentiment analysis has become more and more popular in the field of natural language processing. Multi-modal sentiment analysis mainly concentrates on text, image and audio information. Previous work based on BERT utilizes only text representation to fine-tune BERT, while ignoring the importance of nonverbal information. Most current research methods are fine-tuning models based on BERT that do not optimize BERT's internal structure. Therefore, in this paper, we propose an optimized BERT model that is composed of three modules: the Hierarchical Multi-head Self Attention module realizes the hierarchical extraction process of the features; the Gate Channel module replaces BERT's original Feed-Forward layer to realize information filtering; the tensor fusion model based on self-attention mechanism utilized to implement the fusion process of different modal features. In CMU-MOSI, a public mult-imodal sentiment analysis dataset, the accuracy and F1-Score were improved by 0.44% and 0.46% compared with the original BERT model using custom fusion. Compared with traditional models, such as LSTM and Transformer, they are improved to a certain extent.

Keywords: multi-head self-attention mechanism; multi-modal sentiment analysis; transformer; tensor fusion network

Citation: Wu, J.; Zhu, T.; Zheng, X.; Wang, C. Multi-Modal Sentiment Analysis Based on Interactive Attention Mechanism. *Appl. Sci.* **2022**, *12*, 8174. https://doi.org/10.3390/app12168174

Academic Editors: Shengzong Zhou and Jingsha He

Received: 19 June 2022
Accepted: 3 August 2022
Published: 16 August 2022

1. Introduction

In daily life, with the emergence of various social media, more and more users like to express their feelings on weibo, TikTok, Twitter and other platforms. The content of expression is also varied, from text and emoticons at the beginning to voice and video messages later. The complexity of emotional information is increasing, and the valuable information contained in the information is also increasing. Mining this emotional characteristic information can not only improve services for the platform but also carry out customized recommendation services according to the audience's liking degree. Likewise, it can monitor and manage public opinion for the government. Therefore, how to process and analyze multi-modal emotional information and design a better multi-modal information processing model is very important.

However, most of the later research methods based on multi-modal sentiment analysis are based on fine-tuning the BERT model, and there are few models that directly optimize BERT. Therefore, in this paper, we propose an optimized model based on BERT, LG-BERT. The following innovations are proposed:

(1) Hierarchical multi-attention mechanism used to realize hierarchical extraction of data features;
(2) Gate Channel used to replace the Feed Forward layer in the BERT model to realize information filtering;
(3) Information interaction between different modes realized through tensor fusion model based on self-attention.

2. Related Work

2.1. Deep Learning

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. Gove and Faytong [1] applied Support Vector Machine (SVM) to deal with emotion classification based on document level. Some researchers use reinforcement learning to deal with affective categorization. Chen et al. [2] use it to calculate the emotional value of words and obtain the emotion of sentences by accumulating the emotional value of words. Some researchers also use reinforcement learning to control information input by the gating mechanism [3]. These methods provide a new way of thinking, but it is difficult to take advantage of reinforcement learning in decision-making.

With the emergence of Transformer [4], BERT [5] and other pre-training models, their 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. The LSTM [6] (Long Short-Term Memory, LSTM) model is used more because of its simple structure and different gating mechanism to control the output of features, thus alleviating the problem of gradient disappearance caused by RNN. Therefore, LSTM is also widely used in data modes with sequence characteristics. Zadeh [7] used LSTM to process multi-modal data (audio, test and vision) and used tensor fusion to realize the fusion of different data features. He, J et al. [8] focused on solving deep learning methods that are restricted by a limited receptive field, inflexibility and difficult generalization problems in hyperspectral image classification. They proposed HSI-BERT, where BERT stands for bidirectional encoder representations from transformers and HSI stands for hyperspectral imagery. The proposed HSI-BERT has a global receptive field that captures the global dependence among pixels regardless of their spatial distance. HSI-BERT is very flexible and enables flexible and dynamic input regions. Furthermore, HSI-BERT has good generalization ability because the jointly trained HSI-BERT can be generalized from regions with different shapes without retraining. Chen et al. [9] proposed a text-mining-based accident causal classification method based on a relational graph convolutional network (R-GCN) and pre-trained BERT. Their method avoided preprocessing, such as stop-word removal and word segmentation, which not only preserved the information of accident investigation reports to the greatest extent but also avoided tedious operations. On the other hand, with the help of R-GCN to process the semantic features obtained by BERT representation, the dependence of BERT retraining on computing resources can be avoided. Shuai et al. [10] focused on improving the BERT model's ability in Chinese tasks to understand phrase-level semantic information and proposed an enhanced BERT based on the average pooling (AP-BERT). The AP-BERT model used an average pooling layer to act on token embedding and reconstructs the model's input embedding, which can effectively improve BERT's application effect in Chinese natural language processing. Experimental data showed that the AP-BERT model was enhanced in the tasks of Chinese text classification, named entity recognition, reading comprehension, and summary generation. AP-BERT could not only improve the application effect of the BERT model in Chinese tasks but also can be well applied to other pre-trained language models. Jiaxuan He et al. [11] proposed a multi-modal fusion BERT that can explore the time-dependent interactions among different modalities. Additionally, prior BERT-based methods tended to train the models with only one optimizer to update the parameters. They have set two optimizers for multi-modal fusion BERT and other components of the model with different learning rates, which enabled the model to attain optimal parameters. The results of experiments on public datasets demonstrate that our model was superior to the baselines and achieves the state-of-the-art. Xinhua Zhu et al. [12] proposed a BERT-based Multi-Semantic Learning (BERT-MSL) model with an aspect-aware enhancement for aspect polarity classification, which followed the Transformer structure in BERT and used lightweight multi-head self-attentions for encoding. Experimental results on five SemEval and Twitter datasets demonstrated that their model improved the stability and robustness

of ABSA and significantly outperformed some of the state-of-the-art models under the BERT Post-Training (BERT-PT) environment.

At the same time, the key part of the data features can be found by a 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 RoBERTa, have achieved good results by using a bidirectional pre-training process and increasing the number of parameters. Regarding BERT and other pre-training models as text feature 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 multi-modal emotion classification.

2.2. Multi-Modal Sentiment Analysis Model

Although the model based on text analysis can extract its features and be used for classification problems, it is difficult to achieve the discrimination ability when short sentences, unclear sentences and especially satirical sentences are encountered. Therefore, multi-modal sentiment analysis has become an innovative field in NLP processing. By combining different modes, data can be more informative. The difficulties of multi-modal emotion analysis and processing lie in the following aspects:

(1) Feature extraction: how to use the model to extract the features of different modes;
(2) Feature fusion: how to fuse the features of different modes together to achieve a cross-modal information interaction effect.

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 is 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 the data. For example, using CNN for image-based data will work well. ALBERT dramatically reduces the number of BERT arguments by sharing parameters at the network layer.

Models for feature fusion can be roughly divided into two types; one is early model fusion, also known as the feature-level fusion process [13]. During model training, features are first fused together. For example, different modal features can be fused in a specified dimension and finally sent into the model for training. The other is late fusion, also known as the decision-level fusion [14]. Different models are used to deal with different features, and the final output results can be fused in the way of maximum value and average value according to the size of these values. For some specific fusion cases, researchers can use the method based on tensor fusion to cross-product the data of different modes and by introducing an extra vector to preserve the features of other modes. CM-BERT uses the mask attention mechanism to achieve the fusion of text and speech features. Introducing the mask mechanism, it gives the minimum weight value to features that do not need attention, for cross-modal information fusion. Some researchers also use multiple self-attention blocks to combine different modes in pairs through the self-attention mechanism [15]. The RoBERTa model [16] is used to train audio data as a dynamic presentation of text features to achieve very good results.

2.3. Fusion Method

Fusion methods can be divided into early fusion (also known as feature-level fusion) and late fusion (also known as decision-level fusion), as shown in Figure 1. The early fusion method usually fuses the features of different modalities before inputting the model [7]. Commonly used methods multiply, add or connect the elements at the same position in each mode together along a specific dimension. Late fusion first uses different models to train different modalities and then merges the output results of multiple models [17]. Late fusion methods mainly use fusion rules to determine the combination strategies of different model output results, such as maximum combination, average combination and other combination methods. Others, such as Zadeh et al. [18], use other fusion methods

that use the form of outer tensor product to retain the characteristics of uni-modality and bi-modality. Furthermore, they reduced the calculation by degrading the tensor fusion matrix into a low-rank factor to achieve an efficient feature fusion effect. 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, pre-training models have been widely used in multi-modal emotion classification tasks. A strong pre-training model has a lot of room for improvement. Devlin et al. [5] worked out that the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5%. Wang and Cho [19] showed that BERT is a Markov random field language model. Formulating BERT in this way gives rise to a practical algorithm for generating from BERT that produces diverse and fairly fluent generations. The power of this framework is in allowing the principled application of Gibbs sampling and potentially other MCMC algorithms to be generated by BERT.

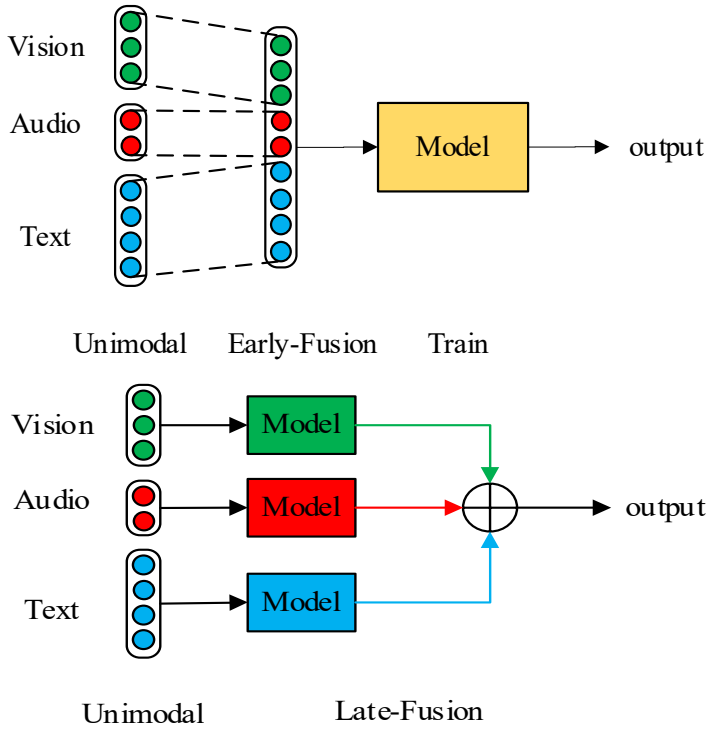

Figure 1. Early Fusion (**upper**) and Late Fusion (**lower**).

3. Interactive Attention Mechanism Based on BERT Model

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. Figure 2 shows the structure of a traditional Attention Mechanism based on LSTM. When processing audio data, in this paper, we use a stacked two-layer LSTM + Full Connection layer (FC) + tanh activation function to extract the audio features. Experiments have found that compared to traditional models, such as GRU, the effect of LSTM is better after obtaining the features, and then the feature fusion method based on TFN is used.

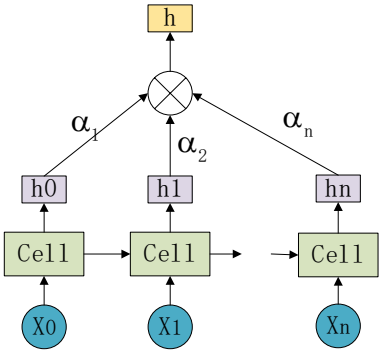

Figure 2. Structure of Attention Mechanism based on LSTM.

However, similar 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 data varies during self-attention. The extraction ability of data features is different for different layers. Therefore, a variable number of self-attention headers is proposed, as shown in Figure 3.

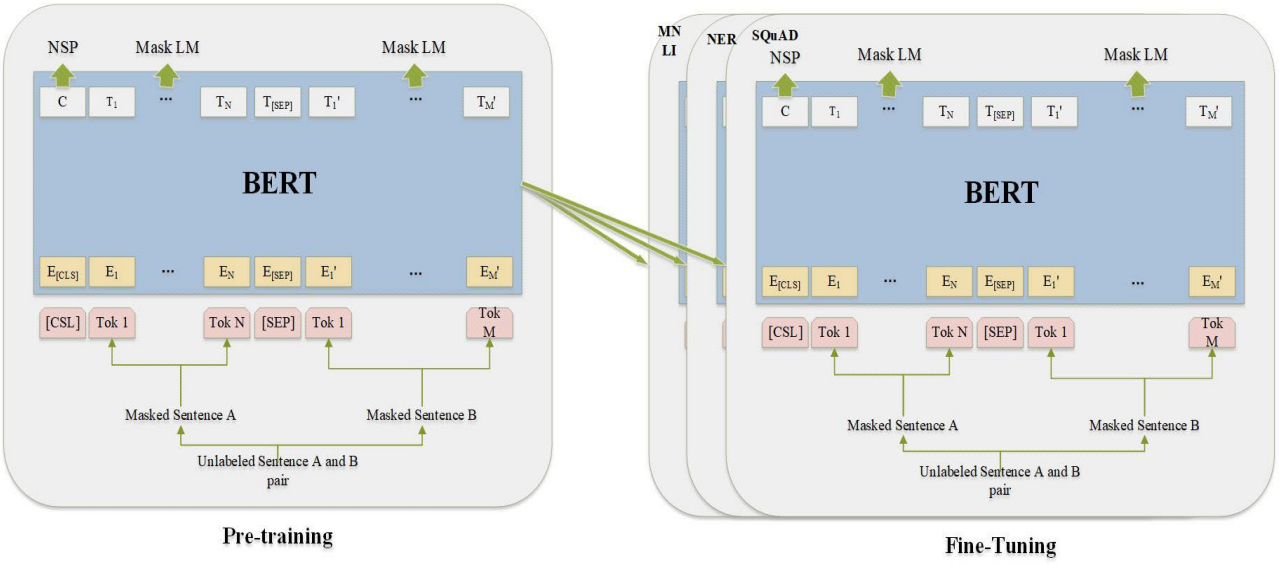

Figure 3. Pre-training and Fine-tunning Procedures for BERT Model.

3.2. Gated Information Channel

The Feed Forward layer in BERT Encoder is essentially a fully connected layer, which expands the embedded feature dimension four times. This is not specified in the Transformer structure. Conversely, the study shows that the fully connected layer may add noise to the original data features. Therefore, a gated information channel is set up to filter out unnecessary data features. Similar to the network structure of GRU [17], this paper designs a similar gated information mechanism. As shown in Figure 4, the gate information channel consists of two parts; one is a memory gate, and the other is an update gate. Memory gates are used to hold valuable information, while additional channels are added to remember new information. The use of another update gate to implement updates to features fully connected layer may add noise to the original data features.

$$G\_M = \sigma(X \otimes M) \quad (1)$$

$$G\_U = \sigma(X \otimes U) \quad (2)$$

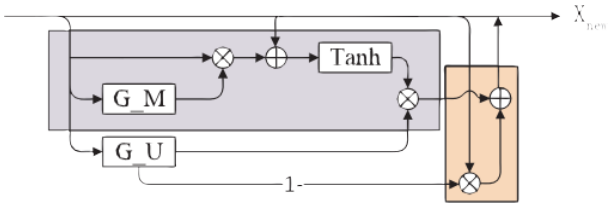

Figure 4. Structure of Gate Channel.

X(X ∈ Rb × len × hn) is the feature after the hierarchical multi-head self-attention mechanism, M and U are the parameter matrices M, U ∈ Rhn × hn. b represents the batch size, len represents the fixed length of text and hn represents the feature dimension.

Memory gates are used to store information and add some new content to features:

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

Next, use the update gate to realize the updating of the original data:

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

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

3.3. Tensor Fusion Method Based on Self-Attention

In this part, as shown in Figure 5, the fusion process of multi-modal data will be introduced. The process can be divided into two stages: the first stage is to use the self-attention mechanism to extract data features; the second stage is the data fusion process in Figure 5. The feature extraction process for text data is as follows: First, the improved LG-BERT is used to process text information, a self-attention mechanism is used to find important parts of the data, and a residual network and BERT layer regularization is used to standardize features. The input data type format of BERT in the classification process is: X = ["CLS", W1, W2, . . . , Wn, "SEP"].

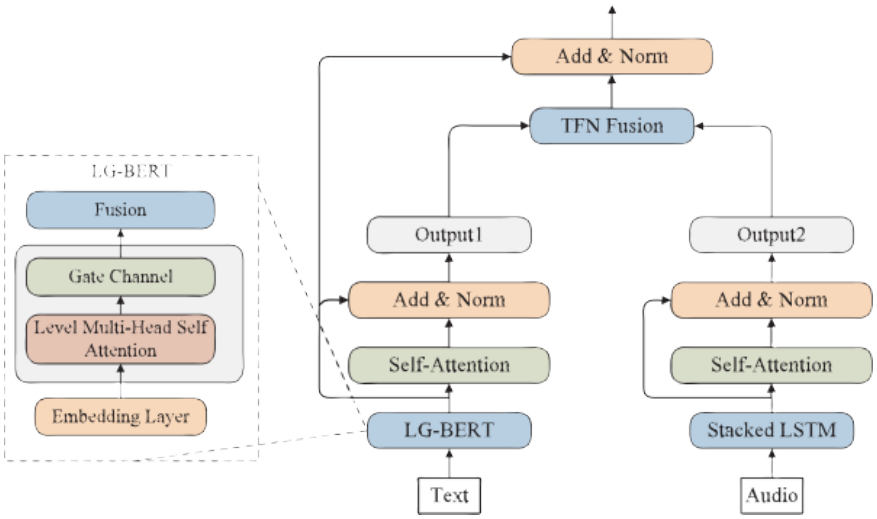

Figure 5. Modal fusion process of text and audio.

Therefore, in Output1, it uses the first item of the output characteristic, which is the characteristic corresponding to the CLS flag bit.

As Figure 6 shows, for feature processing of audio data, we first use a stacked double-layer LSTM to process the audio features. Compared with other traditional models (such as GRU and CNN), LSTM has a better effect.

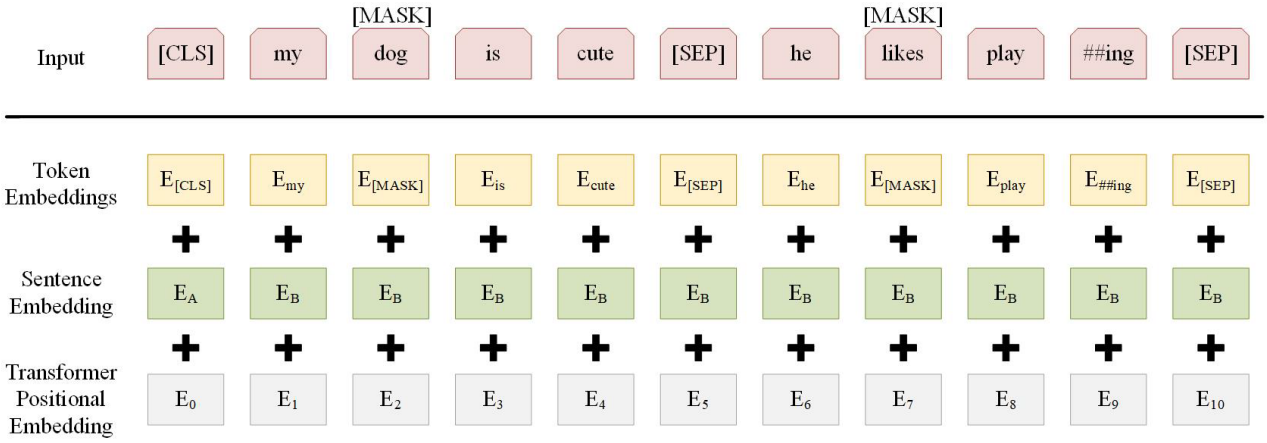

Figure 6. The processing of the BERT Model.

After the processing of 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, the fused features pass through the full connection layer, which is convenient to combine with the original features again.

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

4. Experiment

4.1. Experimental Dataset and Prepossessing

In this part, CMU-MOSI [18] was chosen as the dataset for our experiment. CMU-MOSI consists of 93 videos made up of men and women aged 20–30 (41 were male, and 48 were female). The videos were sliced into 2199 video clips, and each clip was thought to be labeled with a polarity value from −3 to +3. The higher the number, the more polar the emotion. Each video contained an average of 23.2 clips; each clip was 4.2 s in length and contained an average of 12 words.

For the pre-training work of the experiment, the dataset is divided into text and audio. The text belongs to the original data and is processed by built-in functions in BERT. Each sentence is filled and truncated and has a fixed length of dimension 50. Audio data uses processed data. CMU-MOSI provides both raw and processed data characteristics. Here, the CMU-MOSI open dataset processed in the CM-Bert paper is used. Next, the operating environment of the experiment used Pytorch as the Deep learning framework and is coded by Python 3.6.

4.2. Experimental Parameter Configuration

The BERT pre-training model used is the uncased BERT-Base version, consisting of 12-layer Transformer blocks with a hidden layer characteristic of 768 dimensions. The learning rate of the model is 1×10^{-5}, and the optimizer and loss function are AdamW and L1Loss, respectively. The maximum length of the experimental text data is 50, and the characteristic dimension is 768. The embedding dimension of audio data is 5. In the hierarchical multi-head self-attention mechanism, the head number distribution is 16-12-8-4.

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 minimum of 0. In the experiment, the F1-score function in the SK Learn library is used to calculate the F1-score, and the weighting method is adopted.

4.3. Data Processing

The data used in the experiment have been publicly processed, and word embed- ding and alignment about words have been implemented (a word corresponds to a segment of audio and video). The fixed sentence length of a sentence is 50 words. If the sentence length is less than 50, the forward complement 0 is used. If the sentence length exceeds 50, the forward truncation method is used.

5. Analysis of Experimental Results

5.1. Results Analysis of Baseline

The experiment will evaluate the LG-BERT model on the CMU-MOSI dataset; the experimental results are shown in Table 1.

Table 1. The experimental results of different models on the CMU-MOSI dataset, "+" represents the code duplicated by ourselves.

Model	Modality	ACC/%	F1
TFN [5]	T + A + V	77.1	
LMF [6]	T + A + V	76.4	75.7
GME-LSTM [3]	T + A + V	76.5	
MFM [20]	T + A + V	78.1	78.1
RMFN [21]	T + A + V	78.4	78.0
MCTN [22]	T + A + V	79.3	79.1
MulT [23]	T + A + V	83.0	82.8
CM-BERT+	T + A	82.65	82.64
BERT-TA+	T + A	83.38	83.45
LG-BERT (ours)	**T + A**	**83.82**	**83.91**

First, in the processing of multi-modality data based on LSTM, TFN uses the form of 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 N3. LMF degrades the tensor matrix on the basis of the TFN model. Both models use the traditional LSTM model, and the innovation lies in the way of 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, which reinforcement learning is good at, is not introduced, its accuracy is 76.5%. MFM optimizes the joint generation identification model in cross-modal data. Generative factors are shared across channels and contain joint multi-modal 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.

MulT 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. Experiments have found that our model performs better than traditional models. The reason may be that the embedded dimension of audio data is very small, and it is easily affected by noise when passing through the attention module.

Finally, by optimizing the BERT model, the accuracy of the LG-Bert model proposed is improved by 0.44% compared with the BERT model, which provides an idea for further research on the optimization of the BERT model.

5.2. Ablative Study of LG-BERT

Secondly, ablation analysis was conducted for each component in the HG-BERT model. The experimental results are shown in Table 2.

Table 2. Ablation Research of LG-BERT.

Group	Model	Modality	ACC/%	F1/%	Learning Rate
A	BERT	T	83.67	83.70	1×10^{-5}
B1	A + Level Multi-head Attention (LM)	T	82.07	82.08	1×10^{-5}
B2	A + Gate Channel (GC)	T	82.22	82.17	1×10^{-5}
B3	A + Attention Fusion based on TFN (AFT)	T + A	83.38	83.45	1×10^{-5}
B4	A + Fusion (CM-BERT)	T + A	82.36	82.32	1×10^{-5}
D	A + LM + GC + AFT	T + A	83.82	83.91	1×10^{-5}

First, the BERT 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 dual-mode, indicating that the audio data becomes noise during the fusion, which affects the classification effect of the BERT model.

The reasons may lie in two aspects: first, more noise is mixed in audio data extraction and word-based alignment, which affects the accuracy of the audio data. The other is that the processing of audio data using LSTM is not sufficient. Next, 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 paper, the experimental effect of the 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 attention, the attention coefficient obtained from audio and text is used to score with the original text features, to ignore the audio data features. In this paper, the tensor fusion method is used to preserve not only the information of original modes but also the information on interaction between the modes.

6. Conclusions

In this part, we propose an improved model based on BERT. Aiming at the modules that can be further optimized for BERT network structure, several points worth discussing are proposed. First, a hierarchical multi-head self-attention mechanism is used to extract features by using a progressive number of heads, taking advantage of the difference of feature extraction capability of different BERT network layers. Secondly, for the Feed Forward layer of the BERT structure, it is proposed to use the gate channel to replace it and filter the information. Finally, a tensor fusion model based on a self-attention mechanism is proposed to realize feature extraction between cross-modal information by using the self-attention mechanism and the two-mode tensor fusion model.

In future research, the realization of the gated information channel will be further improved. At the same time, it is hoped that the ALBERT model can be used to achieve a significant reduction in the number of BERT model parameters under the condition of ensuring the model performance.

Author Contributions: Conceptualization, J.W.; Data curation, T.Z.; Formal analysis, C.W.; Methodology, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: 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 Hubei University of Technology (HBUT: 2021046).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gove, R.; Faytong, J. Machine Learning and Event-Based Software Testing: Classifiers for Identifying Infeasible GUI Event Sequences. *Adv. Comput.* **2012**, *86*, 109–135.
2. Chen, R.; Zhou, Y.; Zhang, L.; Duan, X. Word-level sentiment analysis with reinforcement learning. *IOP Conf. Series Mater. Sci. Eng.* **2019**, *490*, 062063. [CrossRef]
3. Chen, M.; Wang, S.; Liang, P.P.; Baltrušaitis, T.; Zadeh, A.; Morency, L.-P. Multi-modal sentiment analysis with word-level fusion and reinforcement learning. In Proceedings of the 19th ACM International Conference on Multi-Modal Interaction, Glasgow, UK, 13–17 November 2017; pp. 163–171.
4. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
5. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
6. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
7. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.-P. Tensor fusion network for multimodal sentiment analysis. *arXiv* **2017**, arXiv:1707.07250.
8. He, J.; Zhao, L.; Yang, H.; Zhang, M.; Li, W. HSI-BERT: Hyperspectral Image Classification Using the Bidirectional Encoder Representation From Transformers. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 165–178. [CrossRef]
9. Zail, C.; Huang, K.; Wu, L.; Zhong, Z.; Jiao, Z. Relational Graph Convolutional Network for Text-Mining-Based Accident Causal Classification. *Appl. Sci.* **2022**, *12*, 2482. [CrossRef]
10. Zhao, S.; Zhang, T.; Hu, M.; Chang, W.; You, F. AP-BERT: Enhanced pre-trained model through average pooling. *Appl. Intell.* **2022**. [CrossRef]
11. He, J.; Hu, H. MF-BERT: Multimodal Fusion in Pre-Trained BERT for Sentiment Analysis. *IEEE Signal Process. Lett.* **2021**, *29*, 454–458. [CrossRef]
12. Zhu, X.; Zhu, Y.; Zhang, L.; Chen, Y. A BERT-based multi-semantic learning model with aspect-aware enhancement for aspect polarity classification. *Appl. Intell.* **2022**. [CrossRef]
13. Morency, L.-P.; Mihalcea, R.; Doshi, P. Towards multi-modal sentiment analysis: Harvesting opinions from the web. In Proceedings of the 13th International Conference on Multi-Modal Interfaces, Alicante, Spain, 14–18 November 2011; pp. 169–176.
14. Wang, H.; Meghawat, A.; Morency, L.-P.; Xing, E.P. Select-additive learning: Improving generalization in multi-modal sentiment analysis. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 949–954.
15. Kumar, A.; Vepa, J. Gated mechanism for attention based multi modal sentiment analysis. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 4477–4481.
16. Arjmand, M.; Dousti, M.J.; Moradi, H. Teasel: A transformer-based speech-prefixed language model. *arXiv* **2021**, arXiv:2109.05522.
17. Zhang, S.; Xu, X.; Pang, Y.; Han, J. Multi-layer attention based cnn fortarget-dependent sentiment classification. *Neural Process. Lett.* **2020**, *51*, 2089–2103. [CrossRef]
18. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.-P. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
19. Wang, A.; Cho, K. BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. *arXiv* **2019**, arXiv:1902.04094. [CrossRef]
20. Tsai, Y.-H.H.; Liang, P.P.; Zadeh, A.; Morency, L.-P.; Salakhutdinov, R. Learning factorized multi-modal representations. *arXiv* **2018**, arXiv:1806.06176.
21. Liang, P.P.; Liu, Z.; Zadeh, A.; Morency, L.-P. Multi-modal language analysis with recurrent multistage fusion. *arXiv* **2018**, arXiv:1808.03920.
22. Pham, H.; Liang, P.P.; Manzini, T.; Morency, L.-P.; Póczos, B. Found in translation: Learning robust joint representations by cyclic translations between modalities. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6892–6899.
23. Tsai, Y.-H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multi-modal transformer for unaligned multi-modal language sequences. In Proceedings of the Association for Computational Linguistics Meeting, Florence, Italy, 28 July–2 August 2019; Volume 2019, p. 6558.