Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(10): 18523–18544.
DOI: 10.3934/mbe.2023822
Received: 27 June 2023
Revised: 11 September 2023
Accepted: 13 September 2023
Published: 27 September 2023

Research article

AB-GRU: An attention-based bidirectional GRU model for multimodal sentiment fusion and analysis

Jun Wu[1,2], Xinli Zheng[1], Jiangpeng Wang[1], Junwei Wu[1] and Ji Wang[1,*]

[1] School of Computer Science, Hubei University of Technology, Wuhan 430000, China
[2] Wuhan University of Technology, Wuhan 430000, China

* **Correspondence:** Email: 20061057@hbut.edu.cn; Tel: +8602759750444.

Abstract: Multimodal sentiment analysis is an important area of artificial intelligence. It integrates multiple modalities such as text, audio, video and image into a compact multimodal representation and obtains sentiment information from them. In this paper, we improve two modules, i.e., feature extraction and feature fusion, to enhance multimodal sentiment analysis and finally propose an attention-based two-layer bidirectional GRU (AB-GRU, gated recurrent unit) multimodal sentiment analysis method. For the feature extraction module, we use a two-layer bidirectional GRU network and connect two layers of attention mechanisms to enhance the extraction of important information. The feature fusion part uses low-rank multimodal fusion, which can reduce the multimodal data dimensionality and improve the computational rate and accuracy. The experimental results demonstrate that the AB-GRU model can achieve 80.9% accuracy on the CMU-MOSI dataset, which exceeds the same model type by at least 2.5%. The AB-GRU model also possesses a strong generalization capability and solid robustness.

Keywords: attention mechanism; GRU; multi-modal fusion; multimedia sentiment analysis

1. Introduction

Artificial intelligence is one of the focuses of Internet technology development in recent years, which enables computers to imitate human behavior more intelligently. Among them, sentiment analysis is a technical difficulty that artificial intelligence urgently needs to overcome. It allows computers to cross the dimension of machines and more closely resemble human thinking patterns. Sentiment analysis profoundly explains the development prospect of human-computer interaction and opens the way forward for information technology in the new era. However, going from simple 01-computing to complex and variable brain thinking takes work. Common sentiment analysis methods are built on text data because the textual content reflects the emotional value well. Some

scholars have used parallel convolutional neural networks [1] (CNN) and recurrent neural networks [2] (RNN) for sentiment analysis of the text. Wang et al. proposed an iterative algorithm called SentiDiff to predict sentiment polarities expressed in Twitter messages [3]. Hassonah et al. offered a hybrid machine learning approach to enhance sentiment analysis [4].

However, a single text modality can no longer provide complete data information in the face of complex data types. On social platforms such as Weibo, Twitter and friend circle, people share a large amount of information, such as text, expressions, images, audio and video, to express their emotions in multiple ways. This information also provides a rich database for multimodal sentiment analysis. Audio data contains information such as the size of the voice and the tone of voice. In contrast, image data contains information such as the facial expressions of people and the color tones of images, all of which can assist text content in expressing human emotions better and improving the accuracy of computer judgment of emotional polarity. Huddar et al. presented a novel attention-based multimodal contextual fusion strategy that extracts contextual information among the utterances before fusion [5]. Jiang et al. proposed a model that uses an interactive information fusion mechanism to interactively learn the visual-specific and textual-specific visual representations [6]. Zadeh et al. proposed the multi-attention recurrent network (MARN) [7], novel neural structure for understanding human communication using multi-headed attention modules and mixed long- and short-term memory networks.

Multimodal sentiment analysis methods still have many problems. Regarding feature extraction, the commonly used reinforcement learning methods are often based on the level of words, which ignores the information interaction between words. Most neural network models are also based on single-layer LSTM (long short term memory), which is challenging to extract deeply complex data. Regarding feature fusion, there is a heterogeneous divide between different modalities and the data is in other distribution spaces, which is difficult to measure directly. Existing models are often limited by the exponential growth in computational and memory costs associated with using tensor representations and it is challenging to extend fusion to multiple modalities while maintaining a reasonable model complexity. This paper proposes an attention-based two-layer bidirectional GRU (AB-GRU) multimodal sentiment analysis method to solve the above problems.

First, we preprocess the data of text, audio and image modes, and input the processed data vectors into the double-layer bidirectional GRU network respectively for feature extraction. Then we input the data vector into the attention module to extract the important information and perform the single mode feature fusion. Secondly, we input the extracted multi-modal feature vectors into the low-rank multimodal fusion (LMF) [8] and carry out feature fusion. We add vector 1 to the eigenvectors of three different modes, and align the three modal vectors to form a three-dimensional Cartesian product model, which is then mapped back to the low-dimensional output vector. Finally, we map the results into the sample space to get an output of affective polarity.

This paper aims to learn human emotion polarity using feature extraction techniques and multimodal fusion techniques to learn how emotions are expressed from features such as the content of the text, the priority of audio and human facial expressions in images to achieve more efficient and accurate emotion analysis. The main contributions of this paper are as follows:

1) A two-layer bidirectional GRU network is used to extract multimodal features, which can effectively learn the association of ordered data like text and audio on time series and improve the accuracy of feature extraction. GRU can streamline the gating mechanism and enhance learning

efficiency.

2) Connecting two attention mechanisms after a two-layer bidirectional GRU network can capture important information in the feature vector and enhance the learning efficiency of modal features.

3) LMF converts multiple inputs into a high-dimensional tensor and maps them back to a low-dimensional vector, which can effectively improve the efficiency of operations. Different modalities are decoupled from each other so that the model can extend to data with an arbitrary number of modalities.

2. Related works

2.1. Multimodal fusion

The world comprises countless complex and varied elements and humans can perceive them through sight, hearing, smell, taste and touch to obtain rich knowledge and information. With the development of Internet technology, scholars are also working on making computers learn to imitate this unique way of human information reception, which is also the research direction of artificial intelligence. The research in this direction has obtained excellent results and has been successfully applied in many fields, such as natural language processing, image recognition, recommendation systems and target detection.

Combining the unimodal learning algorithms and techniques of artificial intelligence in different fields, scholars have opened the research on multimodal fusion methods. Multimodal fusion [9, 10] aims to understand and process several different kinds of modal information, including different modalities such as text, audio, image and video, by machine learning to achieve the task of prediction or classification. In the process of data processing by machines, the data of a single modality usually cannot contain complete information and it is challenging for the learning of a single modality to achieve accurate prediction or quickly produce local optimal solutions, so multiple modal data are introduced for fusion learning to improve the learning efficiency. The basic principle of multimodal fusion is the fusion of features of different modal data, i.e., the features of the input data are extracted first. Fusion methods fuse the extracted features of different modalities. Finally, the fused features are input into models such as classification or prediction according to the requirements to obtain the output results. As shown in Figure 1, multimodal fusion methods are divided into early, late and hybrid fusion methods according to the fusion time.

There have been more mature research results on multimodal fusion techniques for different needs. For example, Radford et al. proposed the CLIP [11] model, whose structure consists mainly of a text encoder and an image encoder, which is matched by calculating the similarity between text and image vectors. Zadeh et al. proposed the tensor fusion model (TFN) [12], which uses unimodal features as input and the modal embedding of the 3-fold Cartesian product display of simulated unimodal, bimodal, and trimodal interactions. Memory fusion network (MFN) [13] gives each view an LSTM function component and encodes it independently to send the interactions across views by temporal information.

The use of multimodal fusion techniques for sentiment analysis is also the focus of this research paper. Textual content usually expresses human emotions directly but not comprehensively. Human language is very complex. For example, irony, mockery, rhetorical questions and other emotionally contradictory statements are complex for computers to understand accurately. We, therefore, resort to audio and image data to assist computers in understanding and classifying emotions. The voice

can reflect whether the speaker is anxious or relaxed, and the tone can reflect whether the speaker is angry or calm. All information is contained in the audio data. Vision data can visually represent people's facial expressions and body movements and even the color shades of photos can reflect the photographer's emotion. This information together forms the database of multimodal technology and achieves a more intelligent and accurate multimodal emotion analysis.

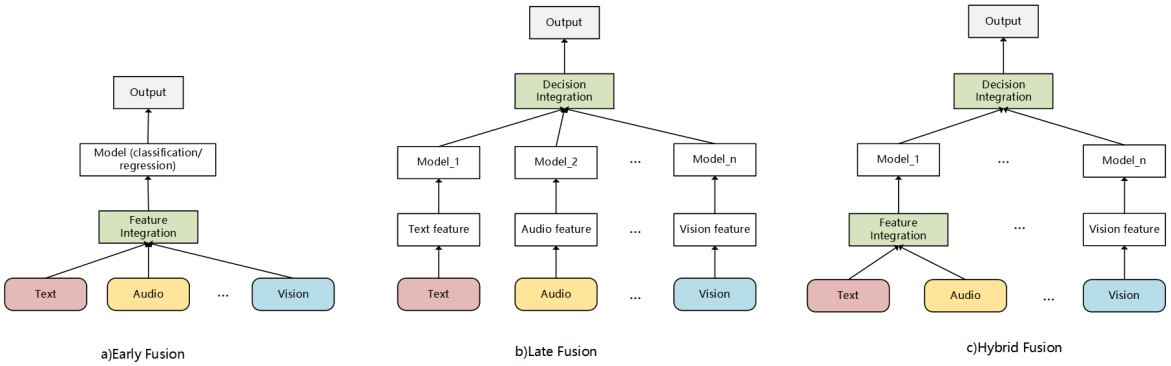

Figure 1. Multimodal fusion methods.

2.2. *Feature extraction*

In multimodal sentiment analysis methods, the effectiveness of feature extraction can directly affect the downstream tasks. Commonly used feature extraction models are CNN [1], RNN [14, 15], LSTM [16, 17] and the current newer transformer [18] and BERT [19, 20]. The sentiment analysis task is mainly based on the text modality. Compared to other modalities, the text modality often contains the richest and most specific sentiment information, while the other modalities play an auxiliary and corrective role. Therefore, among multimodal sentiment analysis methods, the most commonly used feature extraction method is LSTM, which determines the current output by introducing state variables to store past information and current inputs. Also, it solves the problem that RNNs are easily affected by short-term memory, comprehensive sequence information can rarely be kept completely and essential information is easily missed through a unique gating mechanism. Wei et al. proposed a BiLSTM model with multi-polarity orthogonal attention for implicit sentiment analysis [21]. Zhang et al. proposed a recurrent attention LSTM neural network to achieve sentiment analysis by iteratively locating attention regions covering key sentiment words [22].

The GRU [23–25] network used in this paper simplifies the internal structure based on LSTM, simplifies the network model and has fewer model parameters while improves the accuracy rate.

The structure of GRU is shown in Figure 2, which is mainly composed of a reset gate and an update gate.

The role of the reset gate is to determine how much information from the hidden state of the last moment needs to be forgotten and to determine the share of new input and to combine the saved information with the new input.

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right) \tag{2.1}$$

$$\tilde{h_t} = tanh\left(W_{\tilde{h}} \cdot [r_t \times h_{t-1}, x_t]\right) \tag{2.2}$$

where x_t is the current input information, h_{t-1} is the hidden state saved at the last moment, W is the weight, σ is the Sigmoid activation function, compressing the value to between 0 and 1. *tanh* is the tanh activation function compresses the value to between –1 and 1.

r_t is used to adjust the proportion of input information x_t. The value of r_t ranges from 0 to 1, and the smaller the value, the more input information is retained. h_t is the candidate's hidden state. Reset gates help to capture short-term dependencies in the timing information.

Update gates are used to process long-term information, decide how much information from the hidden state of the last moment needs to be remembered and pass the remembered information down the line.

$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right) \tag{2.3}$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \tilde{h}_t \tag{2.4}$$

z_t is used to adjust the degree of history information preservation. z_t takes values from 0 to 1. The smaller the value, the more historical information is preserved. h_t is used to preserve the information that needs to be passed backward.

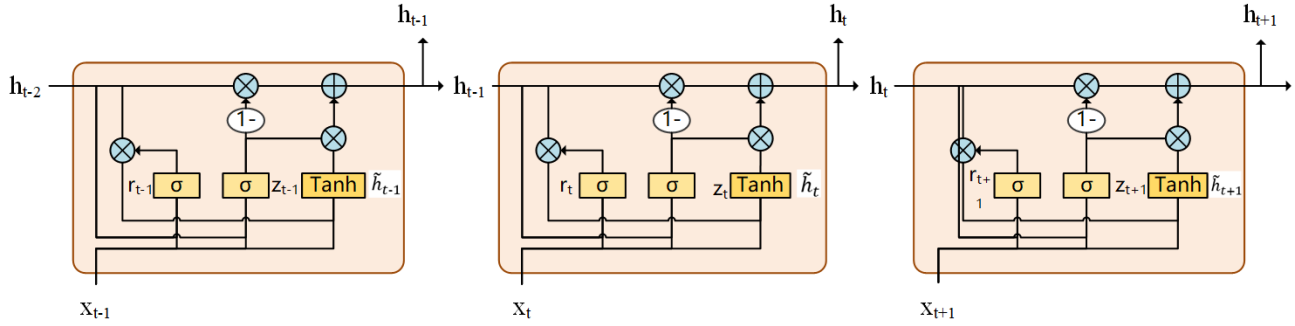

Figure 2. GRU network structure.

The two-layer bidirectional GRU model used in this paper uses a combination of two propagation modes, favorable and negative propagation, based on the ordinary GRU model and a two-layer stacking approach to the bidirectional GRU model. The text and audio modalities are sequential and the previous content will affect the expression of the later content to a certain extent. Thus there is a particular gap between the expressed content of positive and negative propagation. The bidirectional propagation method used in this paper can learn different positive and negative propagation features separately, saving the corresponding hidden information. In the attention module, we combine positive features with hidden positive information and negative features with hidden negative information to focus on the critical information and get more targeted modal features.

For complex modal information, this paper chooses to stack the bidirectional GRU model to achieve higher accuracy feature extraction and improve the overall efficiency of the model because the internal structure of GRU is more straightforward, and the stacking of two layers can also retain its higher computing rate.

3. Attention-based two-layer bidirectional GRU model

3.1. Overall architecture

Different fusion methods fit various tasks. Common multimodal tasks are cross-modal retrieval, sentiment analysis, and audio-visual recognition [26]. CLIP targets cross-modal retrieval tasks, which enables image and text matching. GLCM [27] is a self-supervised method for learning audiovisual representations, which can generalize to both the tasks which require global semantic information and the tasks that require fine-grained spatio-temporal information. In this paper, we investigate the sentiment analysis of multimodal book data [28] and propose an attention-based two-layer bi-directional GRU model, which outperforms most of its current counterparts on sentiment classification tasks.

Compared with existing models, AB-GRU has better classification accuracy and model complexity and can extend to data with an arbitrary number of modalities. Compared with the current outstanding CLIP [11], AB-GRU can better target data with more than two modalities and has low model complexity, few parameters and a high training rate. Compared with traditional TFN [12], AB-GRU uses a stacked GRU network in the feature extraction module and connects two attention layers to enhance the capture of important information. In the feature fusion module, AB-GRU decomposes the weights into low-rank factors to reduce the number of parameters in the model and improve the computation rate. GRU [29] has a wide range of applications in the field of deep learning. For text data and audio data with temporal characteristics, GRU can better learn its features, has a simple structure and a small number of parameters. GRU can greatly improve the operation rate in complex multimodal tasks.

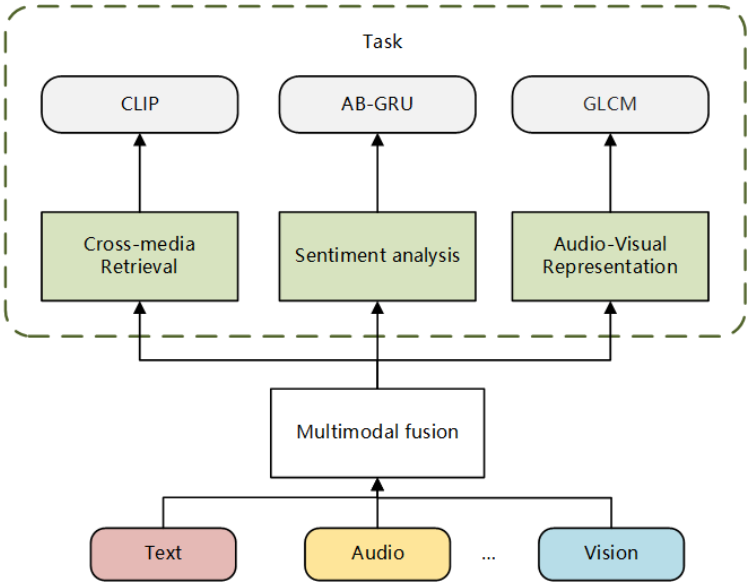

Figure 3. Multimodal fusion methods for different tasks.

The AB-GRU model used in this paper is shown in Figure 3, which consists of a combination of four main modules: input module, feature extraction module, feature fusion module and output module.

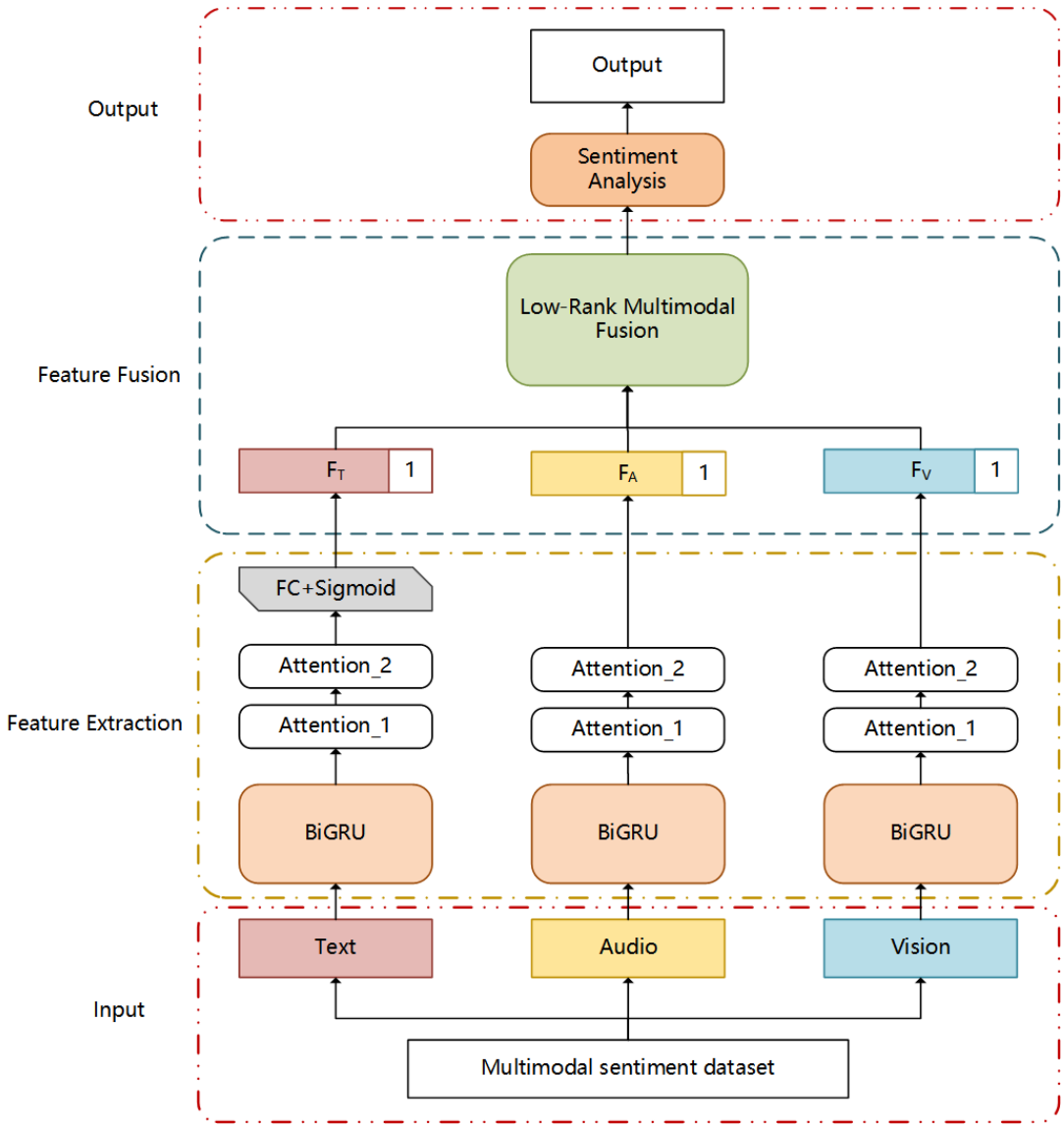

Figure 4. Attention-based two-layer bidirectional GRU multimodal sentiment analysis model.

1) The input module is used to pre-process the multimodal sentiment analysis data, and the data types used in this paper include text, audio and image data.

2) The feature extraction module, which is the focus of improvement in this paper, uses a two-layer bidirectional GRU model based on attention [30] to extract features from the data of the three modalities and obtain the corresponding three feature vectors.

3) The feature fusion module uses LMF for feature fusion to obtain the fused 3D model, which is then mapped back to the low-dimensional output vector.

4) The output module also contains a decision layer. The low-dimensional output vector obtained in the previous step is mapped to the decision layer to obtain the final output by passing the corresponding

single-valued output through the fully connected layer.

3.2. Attention-based two-layer bidirectional GRU model for feature extraction

We preprocess the data of text, audio and image modalities and then use P2FA to perform word alignment to align the three modalities at word granularity and get the data vector of text modality $T = (t_1, t_2, \ldots, t_n)$, n is the vector length of text modality; the data vector of audio modality $A = (a_1, a_2, \ldots, a_m)$, m is the audio modal vector length and the vision modal data vector $V = (v_1, v_2, \ldots, v_l)$, l is the image modal vector length.

The first step in the feature extraction module is to input three modalities, text $T = (t_1, t_2, \ldots, t_n)$, audio $A = (a_1, a_2, \ldots, a_m)$, and vision $V = (v_1, v_2, \ldots, v_l)$, into the attention-based two-layer bidirectional GRU network. Figure 4 shows the feature extraction process of the text.

In the second step, we input the text vector $T = (t_1, t_2, \ldots, t_n)$ into the bidirectional GRU network for learning. The input information will perform update and forget operations in each GRU cell. Then, put the text vector into a second GRU network layer and repeats the above steps. These processes are shown in the second and third modules of Figure 4. Finally, we get the positive hidden layer state ${h_t}^+$ of the text, the negative hidden layer state ${h_t}^-$ and the output $G_T = (G_{t1}, G_{t2}, \ldots, G_{tn})$ of the text after GRU.

Since this paper uses a bidirectional GRU network, output G_T comprises positive and negative propagation processes, so G_T can be decomposed into positive output ${G_T}^+$ and negative output ${G_T}^-$.

Similarly, the audio vector passes through the double-layer bidirectional GRU network to obtain the positive hidden layer state ${h_a}^+$, the negative hidden layer state ${h_a}^-$ and the output $G_A = (G_{a1}, G_{a2}, \ldots, G_{am})$, which can be decomposed into the positive output ${G_A}^+$ and the negative output ${G_A}^-$. The vision vector passes through the two-layer bidirectional GRU network to obtain the positive hidden layer state ${h_v}^+$, the negative hidden layer state ${h_v}^-$ and the output $G_V = (G_{v1}, G_{v2}, \ldots, G_{vl})$, which can be decomposed into the positive output ${G_V}^+$ and the negative output ${G_V}^-$.

In the third step, we put ${h_t}^+$, ${h_t}^-$, ${G_T}^+$ and ${G_T}^-$ into the attention module, the third module in Figure 4. The input content goes through the first layer of attention mechanism $Attention_1$ for unimodal feature fusion: positive hidden features are combined with positive output and negative hidden features with negative output. Then, the attention mechanism learns the critical information of positive and negative directions respectively. Finally, the positive features of the text and the negative features of the text are obtained as follows:

$$ {F_T}^+ = \sum softmax\left[relu\left({h_t}^+ \times {W_T}^+\right)\right] \times \left[relu\left(tanh\left({G_T}^+ \times {W_T}^+\right)\right)\right] \times {G_T}^+ \tag{3.1} $$

$$ {F_T}^- = \sum softmax\left[relu\left({h_t}^- \times {W_T}^-\right)\right] \times \left[relu\left(tanh\left({G_T}^- \times {W_T}^-\right)\right)\right] \times {G_T}^- \tag{3.2} $$

where, ${F_T}^+$ is the positive feature obtained from the text feature vector after $Attention_1$ and ${F_T}^-$ is the negative feature obtained from the text feature vector after $Attention_1$. W_T is the parameter matrix needed to learn. relu and tanh are the activation functions.

Then ${F_T}^+$ and ${F_T}^-$ are input into the second layer of attention mechanism $Attention_2$, to combine the positive and negative features and learn the weights of positive and negative features to obtain the full text features:

$$ F_T = {F_T}^+ \times \theta_T + {F_T}^- \times (1 - \theta_T) \tag{3.3} $$

where F_T is the final text feature obtained from the text vector by the attention-based two-layer bidirectional GRU model and θ_T is the weight needed to learn.

For the audio modality, we combine the audio feature $G_A = (G_{a1}, G_{a2}, \ldots, G_{am})$, the positive output ${G_A}^+$ and the negative output ${G_A}^-$ of the audio feature and the positive hidden layer state ${h_a}^+$ and the negative hidden layer state ${h_a}^-$ of the audio into the first attention mechanism $Attention_1$: Combining the positive hidden feature with the positive output and the negative hidden feature with the negative output. Finally we obtain the positive feature and the negative feature of the audio:

$$ {F_A}^+ = \sum softmax\left[relu\left({h_a}^+ \times {W_A}^+\right)\right] \times \left[relu\left(tanh\left({G_A}^+ \times {W_A}^+\right)\right)\right] \times {G_A}^+ \tag{3.4} $$

$$ {F_A}^- = \sum softmax\left[relu\left({h_a}^- \times {W_A}^-\right)\right] \times \left[relu\left(tanh\left({G_A}^- \times {W_A}^-\right)\right)\right] \times {G_A}^- \tag{3.5} $$

where ${F_A}^+$ is the positive feature obtained from the audio feature vector after $Attention_1$ and ${F_A}^-$ is the negative feature obtained from the audio feature vector after $Attention_1$. W_A is the parameter matrix needed to learn. relu and tanh are the activation functions.

Then ${F_A}^+$ and ${F_A}^-$ are input into the second layer of attention mechanism $Attention_2$, to combine the positive and negative features and learn the weights of positive and negative features to get the complete audio features:

$$ F_A = {F_A}^+ \times \theta_A + {F_A}^- \times (1 - \theta_A) \tag{3.6} $$

where F_A is the final audio feature obtained by passing the audio vector through the attention-based two-layer bidirectional GRU model and θ_A is the weight to be learned.

For the vision modality, we combine the vision feature $G_V = (G_{v1}, G_{v2}, \ldots, G_{vl})$, the positive output ${G_V}^+$ and the negative output ${G_V}^-$ of the vision feature and the positive hidden layer state ${h_v}^+$ and the negative hidden layer state ${h_v}^-$ of the vision into the first attention mechanism $Attention_1$: combining the positive hidden feature with the positive output and the negative hidden feature with the negative output. Finally, we obtain the positive feature and the negative feature of the vision:

$$ {F_V}^+ = \sum softmax\left[relu\left({h_v}^+ \times {W_V}^+\right)\right] \times \left[relu\left(tanh\left({G_V}^+ \times {W_V}^+\right)\right)\right] \times {G_V}^+ \tag{3.7} $$

$$ {F_V}^- = \sum softmax\left[relu\left({h_v}^- \times {W_V}^-\right)\right] \times \left[relu\left(tanh\left({G_V}^- \times {W_V}^-\right)\right)\right] \times {G_V}^- \tag{3.8} $$

where ${F_V}^+$ is the positive feature obtained from the vision feature vector after $Attention_1$, ${F_V}^-$ is the negative feature obtained from the vision feature vector after $Attention_1$, W_V is the parameter matrix to be learned and relu and tanh are the activation functions.

Then ${F_V}^+$ and ${F_V}^-$ are input into the second layer of attention mechanism $Attention_2$, combine the positive and negative features and learn the weights of positive and negative features to obtain the complete vision features:

$$ F_V = {F_V}^+ \times \theta_V + {F_V}^- \times (1 - \theta_V) \tag{3.9} $$

where F_V is the final image feature obtained by passing the image vector through the attention-based two-layer bidirectional GRU model and θ_V is the weight to be learned.

The final feature extraction module gets the outputs: text feature vector $F_T = (F_{t1}, F_{t2}, \ldots, F_{tn})$, audio feature vector $F_A = (F_{a1}, F_{a2}, \ldots, F_{am})$ and vision feature vector $F_V = (F_{v1}, F_{v2}, \ldots, F_{vl})$.

In particular, a fully connected layer is added after the text modality's attention module to reduce the text features' dimensionality. The size of the fully connected layer is the same as the F_T dimension and uses Sigmoid as the activation function.

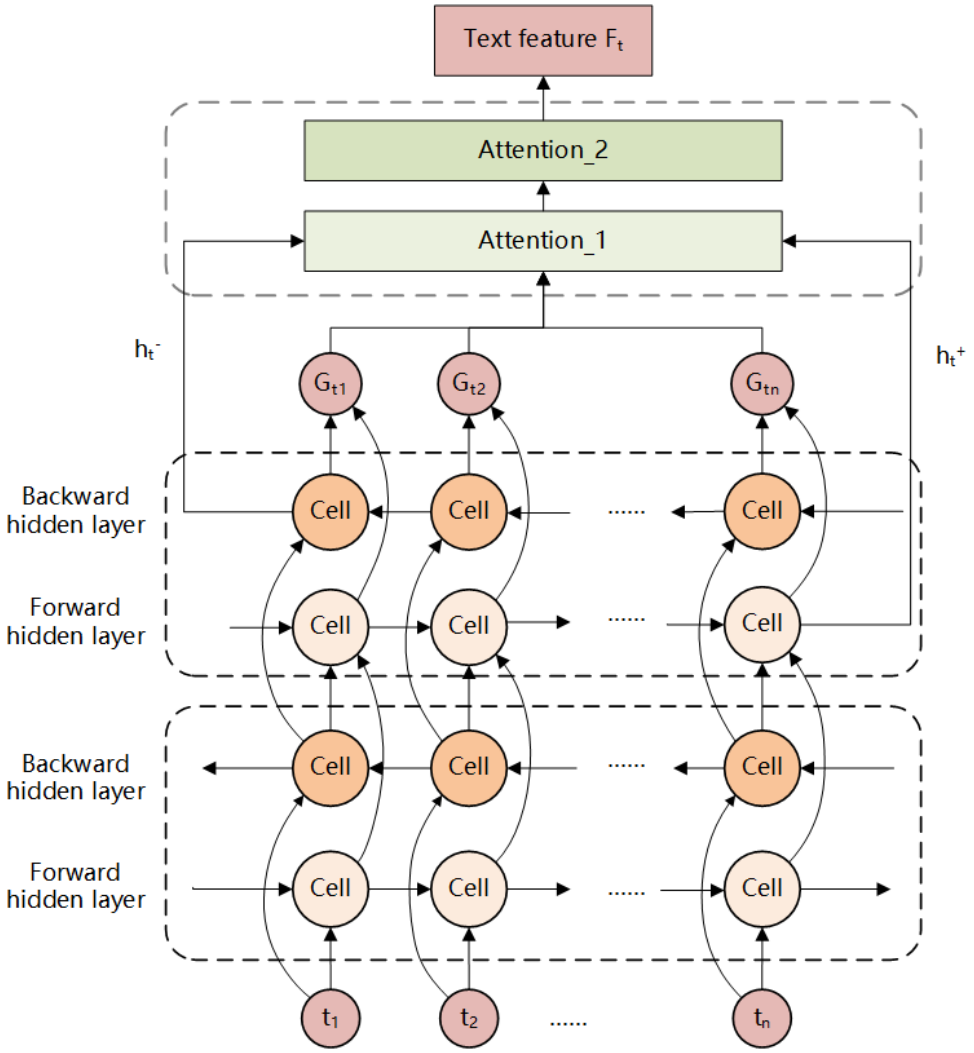

Figure 5. Attention-based two-layer bidirectional GRU model for Text feature extraction.

3.3. Low-Rank multimodal fusion model

The Low-Rank Multimodal Fusion used in this paper in the feature fusion module is a method that uses a low-rank weight tensor to make multimodal fusion efficient without affecting performance. The tensor is powerful in terms of expressiveness and can simulate the alignment and fusion between different modalities very well. The model used in this paper is also an improvement on the tensor fusion model (TFN) [12], which differs from TFN in that LMF decomposes the weights into low-rank factors after the multidimensional model into which the tensor is fused, reducing the number of parameters in the model. Tensor-based fusion can be effectively improved by using parallel decomposition of the low-rank weight tensor and the input tensor to compute tensor-based fusion, which is more efficient than simple splicing or pooling and can scale linearly with the number of modes.

In LMF, multimodal fusion can be described as a multilinear function of:

$$f : D_1 \times D_2 \times \cdots \times D_N \rightarrow H \tag{3.10}$$

where $D_1, D_2, \cdots, D_N$ are the vector spaces of the input modes, N is the number of modes and H is the output vector space.

Multimodal fusion aims to encode the unimodal information of n different modalities and assemble them into a compact multimodal representation. In this paper, we use tensor fusion to store the multimodal interaction information by an additional vector 1 and then obtain a high-dimensional tensor Z_N containing all the modalities by modeling. The tensor is usually obtained by finding the outer product of the input modalities.

$$Z_N = \otimes_{n=1}^{N} z_n \tag{3.11}$$

where $\otimes_{n=1}^{N}$ denotes the tensor outer product of a set of vectors indexed by N. z_n is the input representation of the additional vector 1 for different modalities. As shown in Figure 5, two modalities are aligned by the additional vector 1 to form a tw-dimensional tensor Z_2 , which is then decomposed by the weight tensor W_2 and mapped to a low-dimensional output vector.

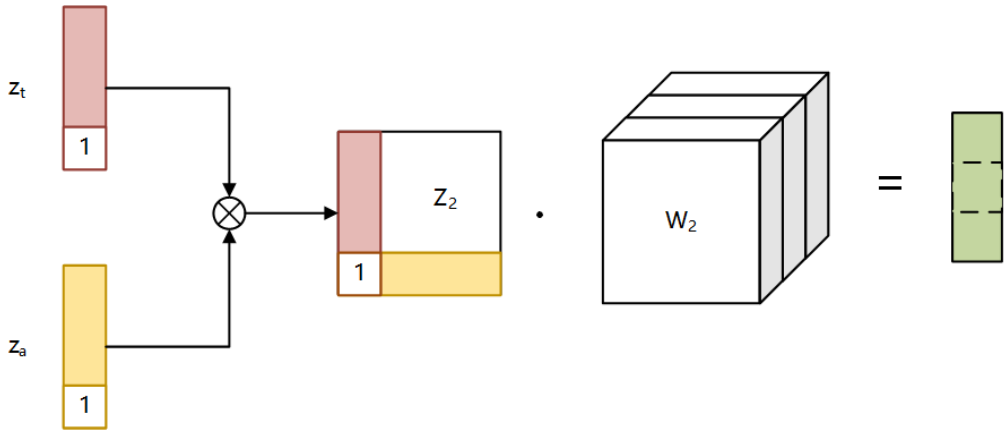

Figure 6. Tensor fusion via tensor outer product.

The feature fusion module in this paper is shown in Figure 6, where the text feature vector $F_T = (F_{t1}, F_{t2}, \ldots, F_{tn})$, the audio feature vector $F_A = (F_{a1}, F_{a2}, \ldots, F_{am})$, and the vision feature vector $F_V = (F_{v1}, F_{v2}, \ldots, F_{vl})$ are input into the low-rank tensor fusion model (LMF) for feature fusion:

A vector with feature value of 1 is appended to each modal feature to store the information interactions between different modalities and to obtain the vector representation Z_T for text features, Z_A for audio features and Z_V for vision features, respectively.

Using the additional vector 1 as the intersection point, we then construct the three modes into a three-dimensional Cartesian product model:

$$Z = Z_T \otimes Z_A \otimes Z_V \tag{3.12}$$

where Z denotes the three-dimensional tensor obtained by fusing the three modalities.

The three-dimensional tensor Z is then mapped back to a low-dimensional vector space to obtain the output of the feature fusion module h:

$$\mathrm{h} = g(Z; W, b) = W \cdot Z + b \tag{3.13}$$

where g(·) is the linear layer function, h is the vector Z generated through the linear layer. W is the weight tensor to be learned and b is the offset.

In LMF, we must map the fused multidimensional tensor back to a low-dimensional output vector to improve the fusion efficiency and facilitate the downstream tasks. In this paper, we parameterize g() as a set of mode-specific low-rank factors for recovering the low-rank weight tensor. By decomposing the weights into a set of low-rank factors and exploiting the nature that the tensor Z can be decomposed into $\{Z_n\}_{n=1}^{N}$, we can compute the output vector h directly, thus reducing the number of parameters involved in the tenderization the computational complexity from N-dimensions to linear levels.

Thus, the vector h can be decomposed as:

$$h = \left(\sum_{i=1}^{r} W_T^{(i)} \otimes W_A^{(i)} \otimes W_V^{(i)}\right) \cdot Z = \left(\sum_{i=1}^{r} W_T^{(i)} \cdot Z_T\right) \circ \left(\sum_{i=1}^{r} W_A^{(i)} \cdot Z_A\right) \circ \left(\sum_{i=1}^{r} W_V^{(i)} \cdot Z_V\right) \tag{3.14}$$

where r is the minimum rank that makes the decomposition valid, W_T is the weight tensor of the text modality, W_A is the weight tensor of the audio modality and W_V is the weight tensor of the vision modality.

Decision classification will be performed in the output module and the output sentiment polarity will be obtained.

We link three fully connected layers and a decision layer after LMF. The size e of the threfully connected layers will be reduced the dimensionality of the vector h layer by layer. We input the vector h obtained from the feature fusion module into the classification module, and reduce its dimensionality through the three fully connected layers. Finally, a single-valued output ρ is obtained. ρ will be input into the decision layer and map to a sample space and the sentiment polarity is positive when $\rho \geq 0$ and negative when $\rho < 0$.

4. Experiments

4.1. Dataset and data processing

In this paper, we use the multimodal sentiment analysis datasets CMU-MOSI [31] and CMU-MOSEI as the experimental datasets. The CMU-MOSI dataset is a collection of 93 opinion videos from YouTube movie reviews, each consisting of multiple opinion clips calibrated by five workers, and finally averaged. The sentiment values for each segment ranged from strongly negative to strongly positive, and the linear scale ranged from –3 to +3. The CMU-MOSEI dataset is the largest multimodal sentiment and emotion recognition dataset available, and contains 23,453 annotated video clips with 250 topics from 1000 different speakers. Each of these video clips contains alignment with audio down to the phoneme level.

Each video is divided into clips based on its transcript. Each paragraph corresponds to the audio and vision of that period to obtain a multimodal sentiment dataset consisting of three modalities: text, audio and vision.

Preprocessing operations are performed for each of the three modalities. The text data are truncated or filled to a length of 50, and word embedding is performed using a 300-dimensional Glove to encode the text sequences into word vector sequences. Enhancement and noise reduction are performed on audio data, and audio features are extracted using the COVAREP acoustic analysis framework. Enhancement and noise reduction are performed on image data using the Facet1 library for extracting visual features.

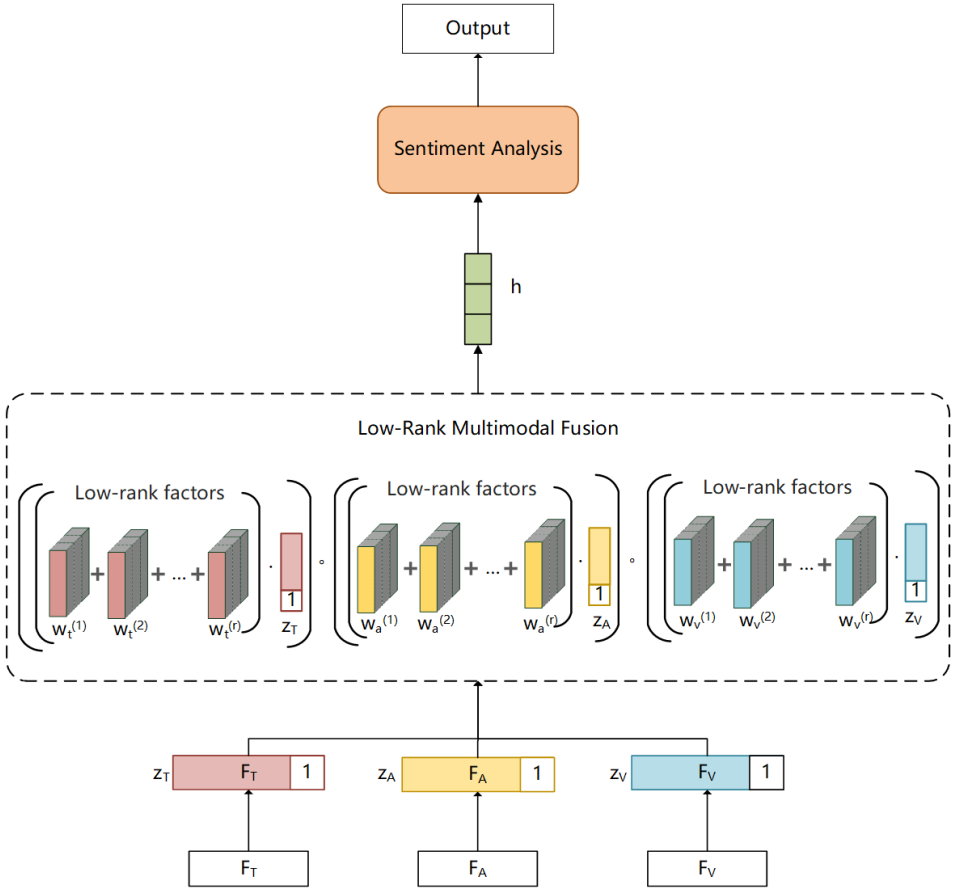

Figure 7. Low-Rank Multimodal Fusion model structure.

4.2. Evaluation standards

In this paper, Accuracy (ACC) and F1-score are used as the evaluation metrics of the model. Accuracy is a primary metric to evaluate the classification task and is the ratio of correct samples to the total number of samples in the classification result:

$$Acc = \frac{n_{correct}}{n_{total}} \tag{4.1}$$

where $n_{correct}$ is the number of correctly classified samples, and n_{total} is the total number.

F1-score is the weighted average of the precision and recall rates:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{4.2}$$

where precision is the accuracy rate and recall is the recall rate. The accuracy rate reflects the ability of the model to distinguish negative samples. The higher the value, the stronger the ability of the model to distinguish negative samples. The recall rate reflects the model's ability to identify positive samples. The higher the value, the stronger the model's ability to identify positive samples. The F1-score is a combination of the two, and the higher the F1-score, the more robust the model.

The F1-score in this paper is calculated by the weighting method. In the baseline of the experiment, if the F1-score has no value, it indicates that the method is not weighted in the calculation.

To enhance the credibility of the experiments, this paper also uses the MAE loss function and the Corr correlation coefficient as the evaluation metrics of the model and the AdamW optimizer as the processor of the network.

$$MAE = \frac{\sum_i \left| y_i - y_i^p \right|}{n} \tag{4.3}$$

where MAE denotes the squared absolute error, y_i denotes the magnitude of the sentiment value of the sample label, y_i^p denotes the magnitude of the predicted value and n denotes the total number of samples.

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var[X]Var[Y]}} \tag{4.4}$$

where Cov(X,Y) is the covariance between X and Y, Var[X] is the variance of X and Var[Y] is the variance of Y.

4.3. Experimental environment and parameter settings

The experimental setting of this paper is shown in Table 1.

Table 1. Experimental environment.

Experimental environment	configuration
Operating system	Windows 10
Processor	Intel(R) Core(TM) i7-10875H CPU @ 2.30GHz
torch	1.9.0cpu
torchvision	0.10.0 + cpu
Programming language	Python3.8
Deep learning framework	Pytorch

In this paper, the loss function used in the experiments is L1Loss, the optimizer is AdamW and the learning rate is 0.001. The activation function used for the text features is Sigmoid, and the activation function for vision features is tanh. The dropout values of the model for the three modes of text, Audio, and Vision are all 0.5. The experimental parameters set on the CMU-MOSI dataset are as follows: the embedding dimensions of text, audio and vision modes are 300, 5 and 20 respectively, and the corresponding hidden dimension in the model is 128, 4 and 16 respectively. Set the batch_size value to 128 and the number of training cycles to 20. Due to the large data set of CMU-MOSEI, we ran the model on GPU and set the experimental parameters as follows: the embedding dimensions of text, audio and vision modes are 300, 35 and 74 respectively, and the corresponding hidden layer dimensions in the model are 128, 16 and 32 respectively. Set the batch_size value to 128 and the number of training cycles to 30.

When feature extraction is performed on the text data, there is an additional fully connected layer for reducing the dimensionality of the text features with a dimension of 128×64. The three fully connected layers after the feature fusion module are used to reduce the dimensionality of the fusion

vectors, which are $(4 + 1) * (16 + 1) * (64 + 1) \times 128$, 128×128 and 128×1. The final single-valued output is obtained.

4.4. Baseline

To validate the performance of the attention-based two-layer bidirectional GRU model proposed in this paper, we compare it with other multimodal fusion models on the CMU-MOSI dataset and CMU-MOSEI dataset.

AB-GRU: This paper proposes the attention-based two-layer bidirectional GRU multimodal sentiment analysis model.

LMF [8]: Low-rank multimodal fusion, which decomposes the weights into low-rank factors, reduces the number of parameters in the model.

TFN [12]: The tensor fusion network is tailored to address the instability of spoken language and accompanying gestures and speech in online videos. It can learn intra-modal and inter-modal dynamics end-to-end.

TFN+: This paper has improved attention-based two-layer bidirectional GRU network by integrating a tensor fusion model.

GME-LSTM [32]: Gated multimodal embedding can solve the fusion challenge when noise is present in the modalities. LSTM with temporal attention can perform word-level fusion with better fusion resolution.

MARN [7]: Multi-attention recurrent network, which discovers interactions between morphologies by using neural components called multi-attention blocks (MAB) and stores them in a mixed memory of recurrent components called long short term hybrid memory (LSTHM).

MFN [13]: Memory fusion network, which explicitly accounts for two interactions in neural structures and models them continuously over time, sends interactions across views with temporal information.

MFM [33]: Multimodal decomposition model optimizes the common generation-discrimination objective across multimodal data and labels by decomposing the representation into two independent sets of factors: multimodal discriminative factors and modality-specific generative factors.

RMFN [34]: Recurrent multi-stage fusion network, which decomposes the fusion problem into multiple stages, each focusing on a subset of multimodal signals, for specialized and efficient fusion.

4.5. Analysis of experimental results

The AB-GRU model used in this paper experimented on the dataset CMU-MOSI, and the results are shown in Figure 8. After the number of training reaches 10, the ACC and loss values of the model gradually become smooth. After several experiments, the final result is 80.9% for ACC and 93.0% for MAE. The current popular multimodal sentiment analysis methods experimented on the dataset CMU-MOSI under the same experimental environment and parameters. The results were compared with the model in this paper, and the results are shown in Table 2. AB-GRU is the attention-based two-layer bidirectional GRU model proposed in this paper. Compared with other multimodal sentiment analysis models, the AB-GRU model showed significant improvements in both ACC and F1 scores, reaching 80.9 and 81.0%, respectively. Compared with the original LMF model, our improved model resulted in a 4.5% increase in classification accuracy. LMF uses LSTM networks for feature extraction, and

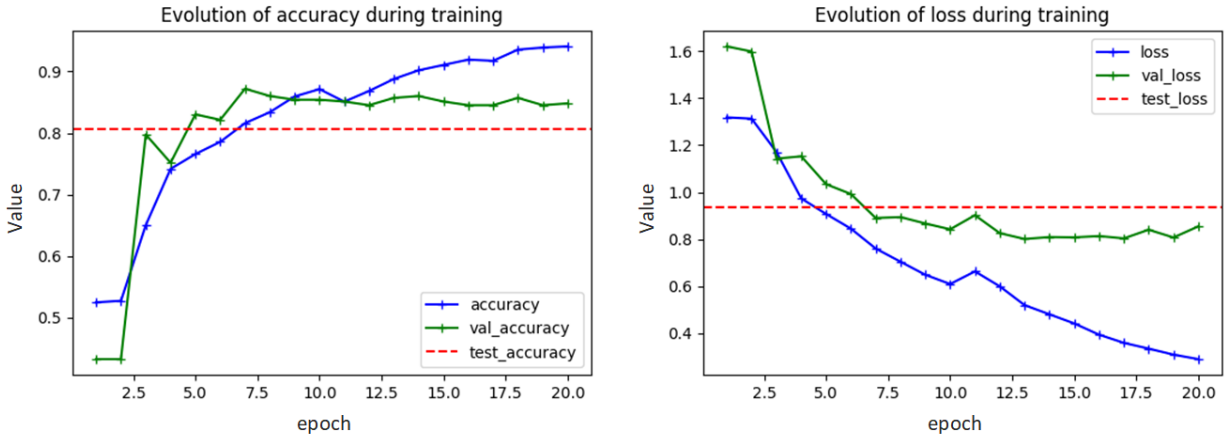

Figure 8. Loss and ACC values of AB-GRU model on CMU-MOSI.

after experiments, it can be seen that using bidirectional GRU networks can improve the efficiency of feature extraction, while stacked GRU networks can effectively improve the accuracy without affecting the experimental rate and focus the vision on different modal data through the attention mechanism important information and extract data features in-depth for downstream fusion tasks.

Table 2. Experimental results of different models on the CMU-MOSI dataset.

Model	ACC/%	F1-score/%	MAE%	Corr%
AB-GRU	80.9	81.0	93.0	65.8
TFN+	80.1	80.1	91.9	69.7
LMF [8]	76.4	75.7	91.2	66.8
TFN [12]	77.1	77.9	95.6	67.2
GME-LSTM [32]	76.5	——	102.0	62.1
MARN [7]	77.1	77.0	96.8	63.2
MFN [13]	77.4	77.3	97.1	62.5
MFM [33]	78.1	78.0	94.5	60.7
RMFN [34]	78.4	78.0	92.9	67.3

Compared with the best RMFN model, AB-GRU improves the training effect by 2.5%. Most current multimodal sentiment classification models focus on modality fusion methods to improve and upgrade. The attention-based bilayer bidirectional GRU model proposed in this paper uses the characteristics of different modal data, which improves and upgrades the feature extraction module and chooses a more suitable low-rank tensor fusion model for feature fusion so that the overall performance has been improved. The high computing rate has been maintained.

The comparative structural analysis of AB-GRU and other models on the CMU-MOSI and CMU-MOSEI datasets is shown in Table 3.

Table 3. Experimental results of different models on the CMU-MOSEI dataset.

Model	CMU-MOSI		CMU-MOSEI	
	ACC/%	F1-score/%	ACC/%	F1-score/%
AB-GRU	80.7	80.9	80.3	80.1
TFN+	80.1	88.0	78.3	78.3
LMF [8]	76.4	75.7	75.2	75.0
TFN [12]	77.1	77.9	76.2	76.1
GME-LSTM [32]	76.5	——	75.6	——
MARN [7]	77.1	77.0	75.9	75.8
MFN [13]	77.4	77.3	76.0	76.0
MFM [33]	78.1	78.0	76.8	76.5
RMFN [34]	78.4	78.0	76.7	76.9

Due to the CMU-MOSEI dataset is more complex, the effect of sentiment analysis are all somewhat weakened, but still it can be seen that the AB-GRU model is superior to other sentiment analysis models.

The experiments show that the AB-GRU model achieves satisfactory performance on both the CMU-MOSI dataset and the CMU-MOSEI dataset. This indicates that the model has good generalization and can adapt to different sentiment analysis tasks and achieve good results on different datasets.

4.6. Ablation experiment

In this section, ablation experiments are set up to verify the importance of different modules in the AB-GRU model. The experimental results are shown in Table 4. We generated Figure 8 from the results in Table 4. The histogram shows that AB-GRU achieves superior results and compares them with the results before improving the different modules in the model.

Table 4. Comparison of ACC and F1-score between different modules.

Model	ACC/%	F1-score/%	MAE%	Corr%
AB-GRU	80.9	81.0	93.0	65.8
TFN+	80.1	80.1	91.9	69.7
single-layer GRU	80.3	79.9	91.8	67.1
LSTM	79.6	79.5	97.0	67.9
Relu	80.6	81.1	93.2	65.7
Add Noisy	80.7	80.7	92.7	67.8

LMF is improved based on TFN, so we combined the AB-GRU model with TFN, whose experimental results are shown in Table 4 for TFN+, which has a significant improvement on TFN and once again verified the effectiveness of the attention-based bilayer bidirectional GRU model on multimodal sentiment classification.

Recurrent neural networks have better results on temporal information such as text and audio. On

choosing LSTM or GRU for feature extraction, we verified that using LSTM combined with LMF for multimodal sentiment classification can achieve an accuracy of 79.6%, which exceeds most similar models but is still lower than the AB-GRU model used in this paper. Second, the stacked bilayer GRU improves the accuracy by 0.6% over the single-layer GRU model and only sacrifices a lower training rate.

We also tested the effect of different activation functions on the performance, and the Sigmoid function finally used was slightly better than using the Relu function. In addition, during the experiments, the text modality has high dimensionality. It is more difficult to process, so it is easy to generate overfitting problems when using the GRU network for feature extraction. We considered adding noise to the text data to improve the model's generalization performance. The results are shown in "noisy" in Table 4. After the experiment, the ACC value did not improve significantly.

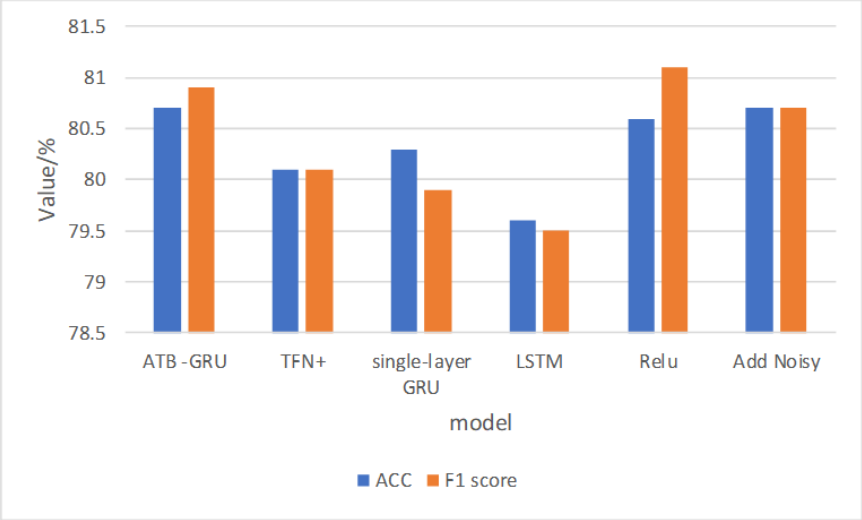

Figure 9. Comparison of ACC and F1-score between different modules.

The different modalities are decoupled from each other in the low-rank tensor fusion model so that it can be extended to data with any number of modalities. To explore the effect of the number and type of modes on the performance, we designed a set of experiments, and the results are shown in Table 5. The training results for different modal combinations following epoch values are shown in Figure 10.

Table 5. Comparison of ACC and F1-score between different modalities.

Model	Modality	ACC/%	F1-score/%
AB-GRU	T+A+V	80.9	81.0
AB-GRU	T+A	79.1	78.9
AB-GRU	T+V	75.9	75.6
AB-GRU	A+V	56.3	56.7

It can be seen from Figure 10 that the classification effect of the A+V combination is significantly lower than the other groups, which shows that in the multimodal emotion classification task, text data

plays a crucial role. In contrast, audio and image data play a supporting role, where audio data is more compatible with text. The audio data is more compatible with the text, and the T + A combination makes the ACC reach 79.1% and the F1-score reach 78.9%, which can handle the sentiment classification task well. The addition of images has further improved the accuracy. However, the improvement is slight but the small improvement plays a crucial role in the face of complex and redundant information.

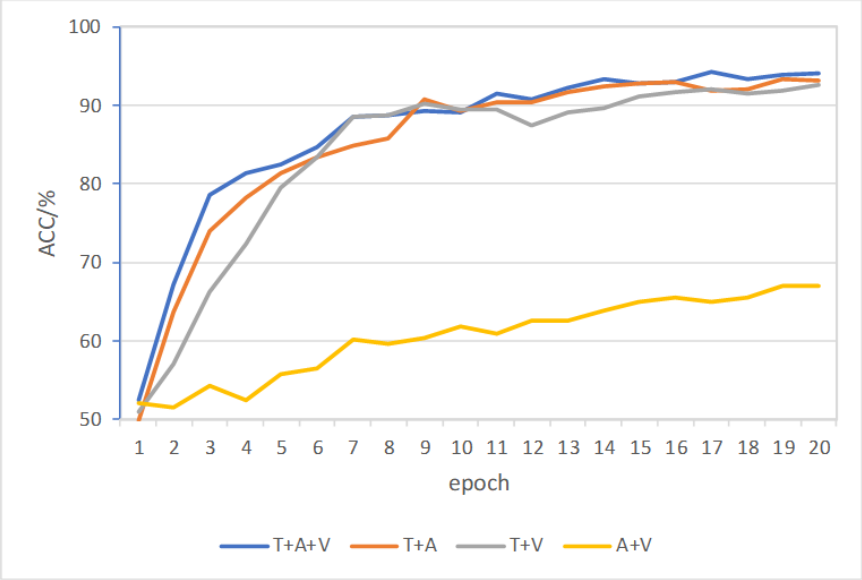

Figure 10. Comparison of ACC between different modalities.

5. Conclusions

In order to solve the problem of the heterogeneinty gap between different modalities and improve the efficiency of feature extraction, this paper proposed an attention-based two-layer bidirectional GRU multimodal sentiment analysis model. The two-layer bidirectional GRU used in this model can effectively learn the text and audio temporal features with a simple structure and fast learning speed. The connected attention layer allows better extraction of essential features. In contrast, the LMF model can reduce the dimensionality of multimodal data, improve the operation rate and increase the accuracy rate. Experimental results show that the performance of the the AB-GRU model proposed in this paper is improved by at least 2.5% compared with other multimodal sentiment analysis models.

In our future work, we will conduct more in-depth research to apply multimodal sentiment analysis methods in different fields. In the medical field, the patient's speech, voice and facial expression can be monitored for condition analysis and timely feedback and treatment can be given. In the short video, classification, integration and recommendation are performed by multimodal methods. Moreover, with the development of technology, we will continue to improve the multimodal sentiment analysis methods, from feature extraction, feature fusion, data pre-processing, and other modules to improve the model's efficiency.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

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), Hubei Provincial Science and Technology Program Project (Grant No.2023BCB041), and The Fundamental Research Funds for the Research Fund of Hubei University of Technology (HBUT: 2021046,21060,21066)

Conflict of interest

The authors declare there is no conflict of interest.

References

1. T. Uyen Tran, H. H. T. Thanh, P. H. Dang, M. Riveill, Multitask apect_based sentiment analysis with integrated bidirectional LSTM & CNN model, in *Proceedings of the 4th International Conference on FutureNetworks and Distributed Systems (ICFNDS)*, (2020), 1–7. https://doi.org/10.1145/3440749.3442656
2. A. Agarwal, P. Dey, S. Kumar, Sentiment analysis using modified GRU, in *Proceedings of the 2022 Fourteenth International Conference on Contemporary Computing*, (2022), 356–361. https://doi.org/10.1145/3549206.3549270
3. L. Wang, J. Niu, S. Yu, SentiDiff: Combining textual information and sentiment diffusion patterns for twitter sentiment analysis, *IEEE Trans. Knowl. Data Eng.*, **32** (2020), 2026–2039. https://doi.org/10.1109/TKDE.2019.2913641
4. M. A. Hassonah, R. Al-Sayyed, A. Rodan, A. M. Al-Zoubi, I. Aljarah, H. Faris, An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter, *Knowl. Based Syst.*, **192** (2020), 105353. https://doi.org/10.1016/j.knosys.2019.105353
5. M. G. Huddar, S. S. Sannakki, V. S. Rajpurohit, Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM, *Multim. Tools Appl.*, **80** (2021), 13059–13076. https://doi.org/10.1007/s11042-020-10285-x
6. T. Jiang, J. Wang, Z. Liu, Y. Ling, Fusion-extraction network for multimodal sentiment analysis, in *Proceedings of the Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference*,(2020), 785–797. https://doi.org/10.1007/978-3-030-47436-2_59
7. A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, L. P. Morency, Multi-attention recurrent network for human communication comprehension, preprint, arXiv: 1802.00923.
8. Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, L. P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, preprint, arXiv: 1806.00064.

9. L. N. Zúñiga-Morales, J. Á. González-Ordiano, J. E. Quiroz-Ibarra, S. J. Simske, Impact evaluation of multimodal information on sentiment analysis. in *Proceedings of the Advances in Computational Intelligence: 21st Mexican International Conference on Artificial Intelligence*, (2022), 18–29. https://doi.org/10.1007/978-3-031-19496-2_2
10. D. Zeng, Y. Yu, K. Oyama, Deep triplet neural networks with cluster-CCA for audio-visual cross-modal retrieval, in *ACM Transaction on Multimedia Computing Communication and Applications (TOMCCAP)*, (2020), 1–23. https://doi.org/10.1145/3387164
11. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., Learning transferable visual models from natural language supervision, preprint, arXiv: 2103.00020.
12. A. Zadeh, M. Chen, S. Poria, E. Cambria, L. P. Morency, Tensor fusion network for multimodal sentiment analysis, preprint, arXiv: 1707.07250.
13. A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, L. P. Morency, Memory fusion network for multi-view sequential learning, preprint, arXiv: 1802.00927.
14. G. Van Houdt, C. Mosquera, G. Nápoles, A review on the long short-term memory model, *Artif. Intell. Rev.*, **53** (2020), 5929–5955. https://doi.org/10.1007/s10462-020-09838-1
15. A. P. Rodrigues, R. Fernandes, A. Shetty, K. Lakshmanna, R. M. Shafi, Real-time twitter spam detection and sentiment analysis using machine learning and deep learning techniques, *Comput. Intell. Neurosci.*, (2022). https://doi.org/10.1155/2022/5211949
16. A. Londhe, P. V. R. D. P. Rao, Aspect based sentiment analysis–an incremental model learning approach using LSTM-RNN, in *Proceedings of the Advances in Computing and Data Sciences: 5th International Conference*, (2021), 677–689. https://doi.org/10.1007/978-3-030-81462-5_59
17. H. Jelodar, Y. Wang, R. Orji, S. Huang, Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach, *IEEE J. Biomed. Health Inform.*, **24** (2020), 2733–2742. https://doi.org/10.1109/JBHI.2020.3001216
18. F. Wang, S. Tian, L. Yu, J. Liu, J. Wang, K. Li, et al., TEDT: Transformer-based encoding-decoding translation network for multimodal sentiment analys, *Cogn. Comput.*, (2022), 1–15 https://doi.org/10.1007/s12559-022-10073-9
19. J. Wu, T. Zhu, J. Zhu, T. Li, C. Wang, A Optimized BERT for Multimodal Sentiment Analysis, *ACM Trans. Multim. Comput. Commun. Appl.*, **19** (2023), 1–12. https://doi.org/10.1080/09540091.2022.2155614
20. A. Bello, S. C. Ng, M. F. Leung, A BERT framework to sentiment analysis of tweets, *Sensors*, **23** (2023), 506. https://doi.org/10.3390/s23010506
21. J. Wei, J. Liao, Z. Yang, S. Wang, Q. Zhao, BiLSTM with multi-polarity orthogonal attention for implicit sentiment analysis, *Neurocomputing*, **383** (2020), 165–173. https://doi.org/10.1016/j.neucom.2019.11.054
22. Y. Zhang, J. Wang, X. Zhang, Conciseness is better: Recurrent attention LSTM model for document-level sentiment analysis, *Neurocomputing*, **462** (2021), 101–112. https://doi.org/10.1016/j.neucom.2021.07.072

23. J. Hassan, U. Shoaib, Multi-class review rating classification using deep recurrent neural network, *Neural Process. Letters*, **51** (2020), 1031–1048. https://doi.org/10.1007/s11063-019-10125-6
24. A. Zouzou, I. E. Azami, Text sentiment analysis with CNN & GRU model using GloVe, in *Proceedings of the 2021 Fifth International Conference On Intelligent Computing in Data Sciences*, (2021), 1–5. https://doi.org/10.1109/ICDS53782.2021.9626715
25. A. G. Eker, K. Eker, N. Duru, Multi-class sentiment analysis from turkish tweets with RNN, in *Proceedings of the 2021 6th International Conference on Computer Science and Engineering (UBMK)*, (2021), 560–564. https://doi.org/10.1109/UBMK52708.2021.9558958
26. L. Zhu, Z. Zhu, C. Zhang, Y. Xu, X. Kong, Multimodal sentiment analysis based on fusion methods: A survey, *Inform. Fusion*, (2023), 306–325. https://doi.org/10.1016/j.inffus.2023.02.028
27. S. Ma, Z. Zeng, D. McDuff, Y. Song, Contrastive self-supervised learning of global-local audio-visual representations, 2021.
28. L. Zhu, M. Xu, Y. Bao, Y. Xu, X. Kong, Deep learning for aspect-based sentiment analysis: A review, *PeerJ Comput. Sci.*, (2022), e1044. https://doi.org/10.7717/peerj-cs.1044
29. X. Liu, J. You, Y. Wu, T. Li, L. Li, Z. Zhang, et al., Attention-based bidirectional GRU networks for efficient HTTPS traffic classification. *Inform. Sci.*, (2020), 297–315. https://doi.org/10.1016/j.ins.2020.05.035
30. J. Wu, T. Zhu, J. Zhu, T. Li, C. Wang, Hierarchical multiples self-attention mechanism for multi-modal analysis, *Multim. Syst.*, (2023). https://doi.org/10.1016/j.ins.2020.05.035
31. A. Zadeh, R. Zellers, E. Pincus, L. P. Morency, Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, preprint, arXiv: 1606.06259.
32. M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, L. P. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning. in *Proceedings of the 19th ACM international conference on multimodal interaction*, (2017), 163–171. https://doi.org/10.1145/3136755.3136801
33. Y. H. H. Tsai, P. P. Liang, A. Zadeh, L. P. Morency, R. Salakhutdinov, Learning factorized multimodal representations, preprint, arXiv: 1806.06176.
34. P. P. Liang, Z. Liu, A. Zadeh, L. P. Morency, Multimodal language analysis with recurrent multistage fusion, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (2018), 150–161. https://doi.org/10.18653/v1/D18-1014

AIMS Press