Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 4767437, 12 pages
https://doi.org/10.1155/2022/4767437

Research Article

Multimodal Sentiment Analysis Based on Cross-Modal Attention and Gated Cyclic Hierarchical Fusion Networks

Zhibang Quan, Tao Sun, Mengli Su, and Jishu Wei

School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China

Correspondence should be addressed to Tao Sun; suntao0906@163.com

Received 21 April 2022; Accepted 15 July 2022; Published 9 August 2022

Academic Editor: Shengrong Gong

Copyright © 2022 Zhibang Quan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Multimodal sentiment analysis has been an active subfield in natural language processing. This makes multimodal sentiment tasks challenging due to the use of different sources for predicting a speaker's sentiment. Previous research has focused on extracting single contextual information within a modality and trying different modality fusion stages to improve prediction accuracy. However, a factor that may lead to poor model performance is that this does not consider the variability between modalities. Furthermore, existing fusion methods tend to extract the representational information of individual modalities before fusion. This ignores the critical role of intermodal interaction information for model prediction. This paper proposes a multimodal sentiment analysis method based on cross-modal attention and gated cyclic hierarchical fusion network MGHF. MGHF is based on the idea of distribution matching, which enables modalities to obtain representational information with a synergistic effect on the overall sentiment orientation in the temporal interaction phase. After that, we designed a gated cyclic hierarchical fusion network that takes text-based acoustic representation, text-based visual representation, and text representation as inputs and eliminates redundant information through a gating mechanism to achieve effective multimodal representation interaction fusion. Our extensive experiments on two publicly available and popular multimodal datasets show that MGHF has significant advantages over previous complex and robust baselines.

1. Introduction

Every day, a large and meaningful amount of information is generated around us. Most of this information is generated on the web, and social media is a centralized area of information on the web. It covers many topics, opinions, sentiments, and emotions closely related to our lives. Multimodal sentiment analysis (MSA) has been an active subfield in natural language processing [1, 2]. This is mainly due to its wide range of applications, such as government elections [3], intelligent healthcare [4], and chatbot recommendation systems for human-computer interaction [5]. Compared to traditional sentiment analysis, MSA uses multiple sources (excerpted raw text, acoustic, and visual information) to make predictions about the sentiment expressed by a specific object in a specific period. One of the multimodal sentiment analysis challenges is to model the interactions between different modalities because they contain supplementary and complementary information [6]. Another factor that limits the performance of multimodal sentiment analysis tasks is data fusion. This is because there are multiple recurring problems, such as missing values and misalignment in visual and auditory modalities [7].

In recent years, researchers have designed sophisticated fusion models. Zadeh et al. [8] designed the tensor fusion network, which uses a Cartesian product to fuse the feature vectors of three modalities; this provided a new idea for multimodal data processing. Tsai et al. [9] designed a multimodal transformer that processed all modalities together to obtain the predicted sentiment scores. Although these methods have achieved good results, a problem that may affect the final prediction effect is that these models ignore the differences between different modalities, which may lead to the loss of crucial prediction information during the modal representation acquisition stage. Hazarika et al. [10] designed a modality-specific and modality-invariant

feature space, combining two types of representations with similarity loss, reconstruction loss, and dissimilarity loss to evaluate the model effect. Yu et al. [11] used a multitask format and introduced an automatic modal label generation module in the training phase to assist the main task channel, saving manual labelling time, and thus improving efficiency. Although these studies also achieved encouraging results, they lacked intermodal information interaction during the modal fusion phase. Doing so may result in the redundant information present in the upper stage being retained in the final prediction stage, making the model performance poor. As shown in Figure 1, there are two opposite prediction results after the same text interacts with different modalities. For example, an ordinary language with ordinary acoustic features is predicted as a negative sentiment. In contrast, the same type of language with positive visual features is predicted as a positive sentiment. This indicates that different modal combinations have a fundamental impact on sentiment prediction. It should be noted that, in Figure 1, "?" indicates that sentiment cannot be accurately identified, "–" represents negative sentiment, and "+" represents positive sentiment. The number of these symbols signifies the intensity of the sentiment.

To address the mentioned issues, inspired by cross-modal matching and interaction modelling, we propose a novel multimodal sentiment analysis framework, MGHF. It includes mid-term interactions performed in the modal representation phase and post-term interactions in the modal fusion phase. This approach allows the model to fully perceive various modalities' potential representational sentiment information, which helps us improve the fusion and prediction results. Although previous studies have shown that text modality is the most critical [9, 12], we still believe that the information implied by any modality should be considered in the MSA task. Specifically, MGHF employs a flexible strategy for modality variability by using appropriate neural networks for different modalities. In the medium-term interaction learning phase, MGHF performs cross-modal attention interactions for acoustic modality, visual modality, and text modality, respectively, to obtain text-based acoustic representation and text-based visual representation. Several past studies [13] have pointed out that task-related information is not evenly distributed across modalities, with the text modality contributing much more than other modalities. There are also studies [8, 9] that would fuse the text-video and audio modalities as a ternary symmetric structure, which does not take into account the variability of the various modalities and thus fails to fuse them correctly. According to previous experience, in order to make the text modality occupy a higher weight than other modalities in the later fusion stage. We combined the text-based acoustic representation with the text representation, the text-based visual representation with the text representation, and the text-based acoustic representation with the text-based visual representation in a two-by-two combination. We also design gated recurrent hierarchical fusion networks that dynamically interact with learning information representations between modal combinations to complement the information between combinations. Our extensive experiments on the publicly available and popular datasets CMU-MOSI [14] and CMU-MOSEI [15] show that MGHF shows strong competitiveness over previous complex interaction and fusion baselines.

The contributions of this paper are summarized as follows:

(i) A gated cyclic hierarchical fusion network for multimodal sentiment analysis is proposed. It dynamically interacts with information representations between 3 different modal pairs. The gated cyclic hierarchical fusion network enables sufficient interaction between each modal pair, eliminates redundant information between modal pairs, and maximizes the retention of valid representations for modal prediction.

(ii) Inspired by distribution matching, we consider the interactions within different modalities. In the modal representation acquisition stage, we make the nonverbal sequences to cross-modal attention with text sequences, which can capture potential representations within different modalities while making the modal representations closer to the real sentiment expressions.

(iii) Experiments conducted on two publicly available multimodal datasets show that our model has significant advantages over previous advanced complex baselines.

2. Related Work

This section introduces multimodal sentiment analysis, as well as related work on multimodal representation learning and data fusion.

2.1. Multimodal Sentiment Analysis. Unlike traditional sentiment analysis, multimodal sentiment analysis often uses multiple sources (excerpted text, audio, video, and other information) to fully and accurately predict the speaker's sentiment orientation. Researchers have various ways to deal with MSA tasks, one of which is representative of the extraction of intramodal temporal information and the other is the extraction of intermodal interaction information. The former mainly uses neural networks such as the Long Short-Term Memory (LSTM) Network [16] for the extraction of modal contextual information [10, 17]. The latter can be further divided into early, late, and hybrid, depending on the fusion stage. Early fusion is the fusion approach used in the pre-extraction phase of the data. Rozgic et al. [18] used early fusion to connect multimodal representations as input to an inference model, which provides a novel idea for modal fusion. Zadeh et al. [19] designed a memory fusion network (MFN) using multiview sequential learning, which explicitly illustrates two interactions in the neural architecture. The post-fusion approach performs a series of necessary processing within the modality and intermodal data fusion in the final stage. Liu et al. [20] proposed a low-rank multimodal fusion approach to reduce the computational

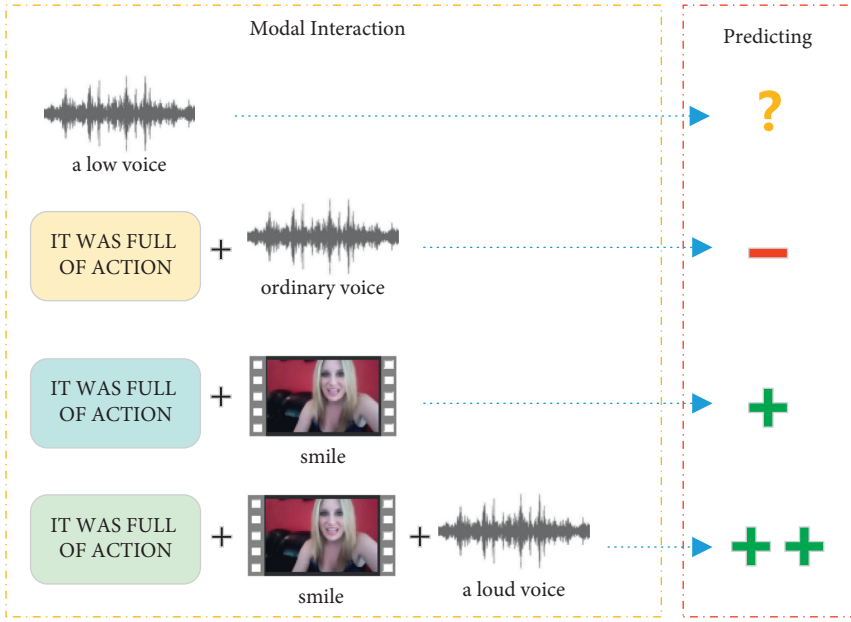

Figure 1: The combination of different modal pairs and sentiment prediction results.

complexity by using low-rank tensor fusion to improve efficiency. Other researchers have used hybrid fusion to improve the performance of MSA tasks. Dai et al. [21] used a simple but very effective hybrid modal fusion approach using weakly supervised multitask learning to improve the generalization performance of the dataset.

We differ fundamentally from previous work in that. First, there is a modal divide between different modalities, and using only the same neural network does not seem to yield useful information. Instead of considering a piece of single contextual information, we use the most appropriate strategy based on the modal sequence characteristics. After obtaining the initial representations, unlike in previous work, our interaction fusion does not only occur in the final stage. Useful potential information can be induced from the companion representations through the intermediate interaction stage. Similarly, the post-interaction stage of the modality is used to better retain information useful for prediction and eliminate redundant information. It is worth noting that instead of the traditional approach of treating text, audio, and video equally, we flexibly utilize the information useful to the task for each modality based on the contribution of the modality.

2.2. Representation Learning and Data Fusion. Representation learning methods can also be applied to multimodal sentiment analysis and have achieved significant results. Wang et al. [22] proposed a recursive attentional change embedding network to generate multimodal shifts. Hazarika et al. [10] proposed a way to learn multimodal invariant and specific representations while combining four different losses to evaluate the performance of the model. Yu et al. [11] proposed self-supervised multitask learning to learn modality-specific representations and introduced a single-peak annotation generation module to assist the main task channel. In the context of sentiment analysis, multimodal fusion is essential because sentiment cues are usually distributed over different modalities [23]. Xiangbo et al. [24] proposed an extended-squeezed-excitation fusion network (ESE-FN) that fuses multimodal features in the modal and channel directions. The network learns extended-squeezed-excitation (ESE) caveats in the modal and channel directions to effectively solve the elderly activity recognition problem. Shu et al. [25] proposed a new weakly shared deep transport network (DTN) for converting cross-domain information from text to images. This provides ideas for interconversion across modalities. Based on this, Tang et al. [26] proposed a new generalized deep transmission network (DTN) for the transmission of information across heterogeneous, textual, and visual domains by establishing parameter sharing and representation sharing layers.

In view of this, our model is based on the late fusion of representation learning. Unlike previous studies, we learn representations across intramodal interactions while employing different combinations of modal interactions to obtain intermodal representations.

3. Materials and Methods

In this section, we will detail the main components of our model and their specific roles.

3.1. Task Setup. Multimodal data sequences in sentiment analysis consist of three main modalities which are the text modality (t), acoustic modality (a), and visual modality (v), respectively. The goal of multimodal sentiment analysis (MSA) is to predict the speaker's emotional polarity from a segment of discourse, which is also the input to the model in this paper. First, given the input discourse $U_{s\in\{t,a,v\}}$, this paper uses U_v to denote visual modal information, U_a to

denote acoustic modal information, and U_t to denote textual modal information. Here, $a \in R^{T_a \times d_a}$, $t \in R^{T_t \times d_t}$, $v \in R^{T_v \times d_v}$, and $T_{s \in \backslash\{t,a,v\backslash\}}$ denote the sequence length of a discourse, and $d_{s \in \backslash\{t,a,v\backslash\}}$ denote the dimensionality of the respective features.

3.2. Overall Architecture. In this paper, our multimodal sentiment analysis architecture consists of three primary and flexible modules as shown in Figure 2. They are the feature extraction module for each modality, the (acoustic-text/visual-text) cross-attention module, and the gated recurrent hierarchical fusion network module. For the text channel, we use pretrained BERT for its high-dimensional semantic extraction. For the acoustic and visual channels, we first feed the initial sequence into a 1D temporal convolution to obtain enough perceptual and temporal information. The obtained (acoustic/visual) representations are then learned cross-modally with textual representations, which can induce potential representational information for both acoustic and visual modalities, synergistic to the overall effective orientation. Notably, this cross-modal matching has been prominent in recent cross-modal learning approaches [27, 28]. Afterward, we feed the output of the two cross-modal attention (text-based acoustic representation and text-based visual representation) and the extracted textual modal representation into a gated recurrent hierarchical fusion network, which eliminates redundant modal information to obtain the final information for prediction. Of course, some of the modules in our model are flexible and can be reconfigured with any suitable baseline to accomplish different types of tasks.

3.3. Modality Representation. The acquisition of representation for our model is divided into three channels, namely, text channel, video channel, and audio channel. In the following, we describe the essential details of the model acquisition of representations.

3.3.1. Text Channel. For the text channel, we fine-tuned the pretrained model BERT [29] used as an extractor of text features, consisting of a 12-layer stacked transformer. The input text is preprocessed and fed to BERT for embedding by adding two special tags CLS and SEP. Consistent with recent work, the first word vector of the last layer is chosen in this paper as the average representation of the representation in the final 768-dimensional implicit state [30].

$$\begin{aligned} t_i &= \{[CLS], w_1, w_2, \ldots, w_n, [\mathrm{SEP}]\}, \\ f_t &= \mathrm{BERT}\left(t_i, \theta_t^{\mathrm{bert}}\right) \in R^{d_t},\ i \in [1, n]. \end{aligned} \tag{1}$$

Here, t represents the initial sequence of text and θ_t^{bert} represents the hyperparameters of the BERT pretrained model.

3.3.2. Audio and Video Channels. For the audio and video channels, we designed two independent modal characterization modules for the nonverbal sequences, and they function before fusion. We followed previous work [11] and processed the raw data using a pretrained toolkit to obtain the initial vector features.

Temporal Convolutions. First, to make our modalities sufficiently perceptible, we pass the input sequence through a one-dimensional temporal convolution layer.

$$U_m^* = \mathrm{Conv1\ } D\left(U_m, k_m\right) \in R^{T_m \times d}, \tag{2}$$

where $\mathrm{Conv1D}(\bullet)$ is the one-dimensional temporal convolution function, k_m is the size of the convolution kernel used by the modality m, U_m is the input sequence of modality m, d is the common dimension, and T_m denotes the discourse length of modality m; here, $m \in \backslash\{a, v\backslash\}$.

Positional Embedding. To equip the sequences with temporal information, following Vaswani et al. [31], the position embedding (PE) is bracketed to U_m^* as follows:

$$U_m^{*'} = U_m^* + \mathrm{PE}\left(T_m, d\right), \tag{3}$$

where $\mathrm{PE}(T_m, d) \in R^{T_m \times d}$, the purpose is to compute the embedding for each position index. $\mathrm{PE}(\bullet)$ represents the position embedding function, $m \in \{a, v\}$.

Cross-Attention Transformers. We then perform cross-modal cross-attention on the resulting sequences, which induces potential representational information for both acoustic and visual modalities that are synergistic to the overall practical orientation. It is worth noting that our cross-modal attention occurs only between text and acoustic modalities and between text and visual modalities, which allows the text modality that contributes most to the task to be weighted higher than the other modalities and ensures the relative independence of the visual and acoustic channels. We justify this approach in Section 5.2.1.

$$\begin{aligned} C\_\mathrm{Attention}_{a-t}(Q, K, V) &= \mathrm{softmax}\left(\frac{Q_t K_a^T}{\sqrt{d_h}}\right) V_a, \\ C\_\mathrm{Attention}_{v-t}(Q, K, V) &= \mathrm{softmax}\left(\frac{Q_t K_v^T}{\sqrt{d_h}}\right) V_v, \end{aligned} \tag{4}$$

where Q_t represents the query vector for the text modality and K_a, V_a, K_v, and V_v denote the key vectors and value vectors of the acoustic and visual modalities. $\mathrm{softmax}(\bullet)$ represents the softmax function, d_h represents the dimensionality of the modality, and T represents transpose.

Transformer computes multiple parallel attentions, and the output of each attention is called a head. The i^{th} head is computed as

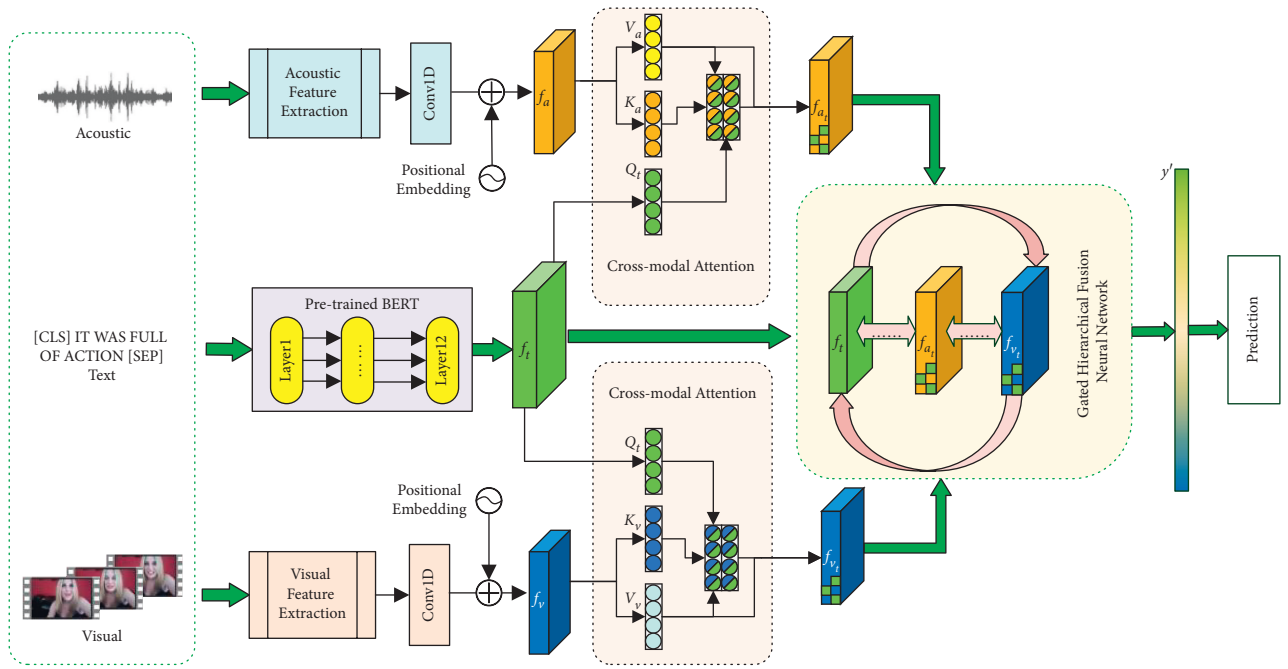

Figure 2: MGHF: cross-modal attention with hierarchical recurrent fusion network.

$$\text{head}_m^i == \text{Attention}_m\left(Q_t W_i^{Q_t}, K_m W_i^{K_m}, V_m W_i^{V_m}\right), \tag{5}$$

where $W_i^{Q_t} \in R^{d_t \times d_q}$ is the weight matrix of Q_t when computing the head of the i^{th} text modality; $W_i^{K_m} \in R^{d_m \times d_k}$ is the weight matrix of K_m when computing the head of the $i^{th}m$ modality; and $W_i^{V_m} \in R^{d_m \times d_v}$ is the weight matrix of V_m when computing the head of the $i^{th}m$ modality, where $m \in \backslash\{a, v\backslash\}$.

After that, we connect all heads of m modalities, which is denoted as Y_m^* as follows:

$$\begin{aligned} Y_m^* &= \text{MultiHead}\left(Q_t, K_m, V_m\right) \\ &= \text{Concat}\left(\text{head}_m^1, \text{head}_m^2, \ldots, \text{head}_m^n\right) W_m^o, \end{aligned} \tag{6}$$

where W_m^O is the weight matrix multiplied after the splicing the head of m modalities and n denotes the number of self-attention heads we use. Here, we have $n = 10$, Concat $(\bullet)$ is the splicing operation, $m \in \backslash\{a, v\backslash\}$.

Thus, the text-based acoustic representation f_{a_t} and the text-based visual representation f_{v_t} can be obtained.

$$\begin{aligned} f_{a_t} &= \text{MultiHead}\left(Y_a^*; \theta_a^{c-att}\right), \\ f_{v_t} &= \text{MultiHead}\left(Y_v^*, \theta_v^{c-att}\right), \end{aligned} \tag{7}$$

where $\theta_a^{att} = \{W_a^Q, W_a^K, W_a^V, W_a^O\}$ and $\theta_v^{att} = \{W_v^Q, W_v^K, W_v^V, W_v^O\}$ represent the main hyperparameters required for the cross-attention module.

3.4. Gated Cyclic Hierarchical Fusion Networks. In previous studies [10, 11], after obtaining valid representations, most of the modal representations are simply spliced directly for final prediction. This can inadvertently add redundant information to them. To allow the redundant information in the representations to be effectively removed, we designed a gated recurrent fusion network (see Figure 3). This module is flexible and can be paired with other benchmarks to enhance the effect. Of course, we also verified the effectiveness of the hierarchical fusion network.

We used the text-based acoustic representation f_{a_t} and text-based visual representation f_{v_t} as well as text representation f_t as inputs to the gated recurrent hierarchical network. Previous experience [9, 12] has shown that the text modality contributes much more to the task than the other modalities. Given this, we combined the text-based visual representation, the text-based acoustic representation, and the text representation in two combinations to ensure that the text modality accounts for a high weight, which would result in three combinations of representations.

$$\begin{aligned} f_{a_t \oplus t} &= \text{Concat}\left(f_t, f_{a_t}\right), \\ f_{v_t \oplus t} &= \text{Concat}\left(f_t, f_{v_t}\right), \\ f_{a_t \oplus v_t} &= \text{Concat}\left(f_{a_t}, f_{v_t}\right). \end{aligned} \tag{8}$$

where Concat $(\bullet)$ denotes the combination operation, $f_{a_t \oplus t}$ denotes the combination of text-based acoustic representation with text, $f_{v_t \oplus t}$ denotes the combination of text-based visual representation with text, and $f_{a_t \oplus v_t}$ denotes the combination of text-based acoustic representation with text-based visual representation.

After obtaining the specified three combinations, we fed them into a bi-directional gated recurrent network (Bi-GRU). The purpose of doing so is to allow the information between different modalities to be fully perceived and to effectively remove redundant and irrelevant information from the representations through the gating mechanism. We also employ a bi-directional long and short memory

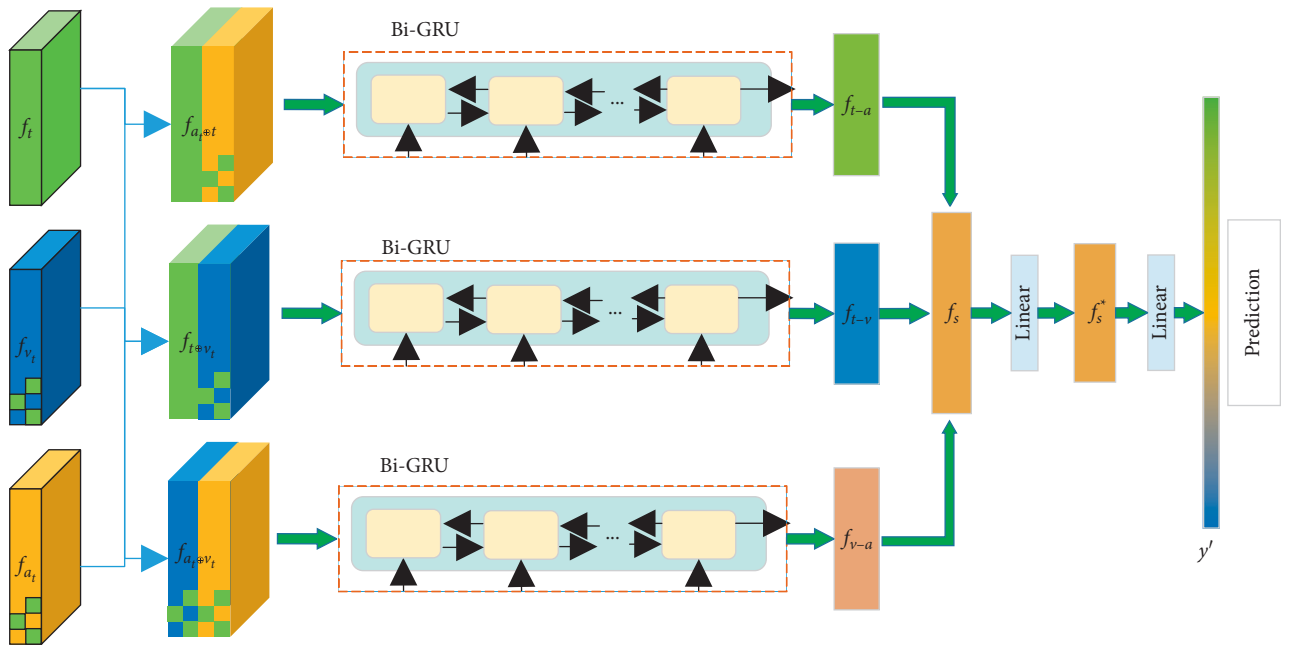

Figure 3: Gated cyclic hierarchical fusion network.

(Bi-LSTM) network. By comparison, we found that the former has more straightforward parameters and faster training speed, and its results are comparable.

$$\begin{aligned} f_{t-a} &= Bi\_GRU\left(f_{a_t \oplus t}, \theta^{gru}\right), \\ f_{t-v} &= Bi\_GRU\left(f_{v_t \oplus t}, \theta^{gru}\right), \\ f_{a-v} &= Bi\_GRU\left(f_{a_t \oplus v_t}, \theta^{gru}\right), \end{aligned} \tag{9}$$

where $Bi\_GRU(\bullet)$ represents the bi-directional gated recurrent cell network and θ^{gru} represents the hyperparameters of the gated recurrent cell network.

After that, we combine the outputs of the gated cyclic hierarchical fusion networks and feed them into the fully connected layer for the final prediction.

$$\begin{aligned} f_s &= \text{concat}\left(f_{t-a}, f_{t-v}, f_{a-v}\right), \\ f_s^* &= \text{ReLU}\left(W_{l1}^{sT} \otimes f_s + b_{l1}^s\right), \end{aligned} \tag{10}$$

where $W_{l1}^s \in R^{(d_t+d_a+d_v)\times d_s}$ and ReLU are the relu activation functions and $\otimes$ represents the elemental product.

Finally, f_s^* is used as the final representation and for the prediction task.

$$y' = \text{ReLU}\left(W_{l2}^{sT} \otimes f_s^* + b_{l2}^s\right), \tag{11}$$

where $W_{l2}^s \in R^{d_s \times 1}$.

4. Experiment

In this section, we will detail the specifics of our experiments.

4.1. Datasets. *CMU-MOSI* [14]. The Multimodal Sentiment Intensity Corpus dataset is a collection of 2199 viewpoint video clips. This dataset is a popular benchmark for multimodal sentiment analysis. Each opinion video is annotated with sentiment in the range of [−3, 3]. The dataset is strictly labelled using tags for subjectivity, emotional intensity, per-frame, per-viewpoint annotated visual features, and per-millisecond annotated audio features.

CMU-MOSEI [15]. The multimodal Opinion Sentiment and Sentiment Intensity dataset is the largest multimodal sentiment analysis and recognition dataset. The dataset is an improved version of the CMU-MOSEI dataset. MOSEI contains more than 23,500 sentence expression videos from more than 1,000 online YouTube speakers. The dataset is gender-balanced. All sentences were randomly selected from different videos of topics and monologues. Videos were transcribed and correctly punctuated. We give the detailed dataset settings in the experiments (see Table 1).

4.2. Modality Processing. To ensure fair competition with other baselines, we follow previous work [11] and treat the three modalities as a typical tensor described as follows:

Text Modality. Most previous studies have used glove [32] as a source of word embedding and achieved good results. Considering the strong performance of pretrained models, we prefer to use the pretrained language model BERT [29]. For a fair and objective comparison, we adopted the latter as the processing tool for our text modality.

Audio Modality. For audio data, the acoustic analysis framework COVAREP [33] was used to extract up to 12 Mel-frequency cepstral coefficients, pitch, turbid/apparent segmentation features, and so on. All features are related to mood and intonation. It is worth noting that acoustic features are processed to align with the text features.

Table 1: MOSI and MOSEI dataset size settings.

Dataset	MOSI	MOSEI
Train	1284	16326
Valid	229	4659
Test	686	1871
All	2199	22856

Video Modality. Video modality raw features are used to extract facial expression features using Facet (https://imotions.com/platform/), which includes facial action units and facial poses based on the Facial Action Coding System (FACS) [34]. The process is repeated for each sampled frame within the vocalized video sequence.

Eventually, we align the initial modalities with the text for the alignment operation. This will allow our experiments to proceed appropriately and ensure fair experimental comparison results.

4.3. Evaluation Metrics. Again, to be fair, we split the MSA task into a regression task and a classification task. This paper will have five valuation metrics, which are: secondary precision (ACC-2) and F1-score. Mean Absolute Error (MAE): it directly calculates the error between the prediction and the authentic number labels. Level 7 Precision (ACC-7) and Pearson Correlation (Corr) measure the standard deviation from the human-annotated actual value. It is worth noting that the secondary precision and F1 scores were divided into two groups: negative and non-negative feelings (including neutral feelings), and negative and positive feelings, respectively. In addition to the value of MAE, higher scores imply better results.

4.4. Baseline. We compared the performance of MGHF with several multimodal fusion frameworks, including state-of-the-art models, as follows.

4.4.1. Previous Models

(i) *TFN*. Tensor fusion network [8] is based on Cartesian product to calculate the tensor of each modality for capturing the interaction information of unimodal, bimodal, and three modalities.

(ii) *LMF*. Low-order multimodal fusion [20] is an improvement of the tensor fusion network (TFN) to reduce the computational complexity and improve the efficiency by using low-order tensor fusion.

(iii) *MFM*. Multimodal Factorization Model [35] demonstrates flexible generation capability by adjusting independent factors and reconstructs missing modes.

(iv) *MULT*. Multimodal Transformer (MULT) [9] extends the multimodal converter architecture using directed pairwise cross-attention, which converts one modality to another using directed pairwise cross-attention.

(v) *ICCN*. Interaction Canonical Correlation Network (ICCN) [13] learns correlations between text, audio, and video through Deep Typical Correlation Analysis (DCCA).

(vi) *MISA*. Learning Modality-Invariant and Modality-Specific Representations (MISA) [10] combines a combination of distribution similarity, orthogonal loss, reconstruction loss, and task prediction loss for learning the representation of different modalities and the representation of fused modalities.

(vii) *MAG-BERT* [36]. A multimodal adaptation gate was designed for the BERT alignment gate and inserted into the general BERT model to optimize the fusion process.

4.4.2. State-of-the-Art. For sentiment analysis tasks, the results of Self-MM [11], a self-supervised multitask learning framework, on both MOSI and MOSEI datasets represent state-of-the-art (SOTA) models. Self-MM assigns a single-peaked training task with automatically generated labels to each modality, allowing multimodal sentiment analysis tasks to be performed in a multitask context.

5. Results and Discussion

In this section, the experimental results of the model are analysed and discussed in detail.

5.1. Quantitative Results. We compared the MGHF with currently popular benchmarks, including the state-of-the-art (SOTA) model (see Tables 2 and 3). For a fair comparison, we divided the models into two categories depending on the data setup, aligned and unaligned. In our experiments, first, compared with the aligned advanced models, our models all achieved similar or even surpassed results. In addition, our models achieve significant gains on all indicators of the regression as well as on some of the categorical indicators compared to the unaligned models. In addition, we reproduce two strong baselines, MISA and self-mm, under the same conditions. We find that MGHF outperforms them on most indicators. On the MOSI dataset, MGHF achieves competitive scores on both classification tasks. On the regression task, MGHF also improves the SOTA model by various degrees. Our model also outperforms some complex fusion mechanisms, such as TFN and LFN. The above results show that our model can be applied to different data scenarios and achieve significant improvements. We visualized some of the metrics, which can help us visualize how the model is performing (see Figure 4).

5.2. Ablation Study. We set up ablation experiments to verify the performance of our model, which is divided into the following main parts.

Table 2: Results on MOSI. Note: (B) Means the language features are based on BERT; model with * represents the best results for recurrence under the same conditions. ○ is from[10], and ◇ is from [11]. In indicators Acc-2 and F1-score, the left side of "/" is calculated for negative and non-negative sentiment, while the right side of "/" is calculated for negative and positive sentiment.

Models	MOSI					Data setting
	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1-score (↑)	
TFN (B)○	0.901	0.698	34.9	−/80.8	−/80.7	Unaligned
LMF (B)○	0.917	0.695	33.2	−/82.5	−/82.4	Unaligned
MFM (B)○	0.877	0.706	35.4	−/81.7	−/81.6	Aligned
MULT*	0.918	0.680	36.47	77.93/79.3	77.91/79.34	Aligned
ICCN (B)◇	0.860	0.710	39.0	−/83.0	−/83.0	Unaligned
MISA (B)◇	0.783	0.761	42.3	81.8/83.4	81.7/83.6	Aligned
MAG-BERT (B)◇	0.731	0.789	—	82.54/84.3	82.59/84.3	Aligned
Self-MM (B)◇	0.713	0.798	—	84.42/85.95	84.42/85.95	Unaligned
MISA (B)*	0.759	0.787	42.57	81.05/82.93	81.03/82.97	Aligned
Self-MM (B)*	0.718	0.796	45.77	83.09/84.09	83.10/84.96	Aligned
MGHF (B)	0.709	0.802	45.19	83.38/85.21	83.32/85.21	Aligned

Table 3: Results on MOSEI. Note: (B) Means the language features are based on BERT; model with * represents the best results for recurrence under the same conditions. ○ is from [10], and ◇ is from [11]. In indicators Acc-2 and F1-score, the left side of "/" is calculated for negative and non-negative sentiment, while the right side of "/" is calculated for negative and positive sentiment.

Models	MOSEI					Data setting
	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1-score (↑)	
TFN (B)○	0.593	0.700	50.2	−/82.5	−/82.1	Unaligned
LMF (B)○	0.623	0.677	48.0	−/82.0	−/82.1	Unaligned
MFM (B)○	0.568	0.717	51.3	−/84.4	−/84.3	Aligned
MULT○	0.580	0.703	51.8	−/82.5	−/82.3	Aligned
ICCN (B)○	0.565	0.713	51.6	−/84.2	−/84.2	Unaligned
MISA (B)◇	0.555	0.756	52.2	83.6/85.5	83.8/85.3	Aligned
MAG-BERT (B)◇	0.539	0.753	—	83.79/85.23	83.74/85.08	Aligned
Self-MM (B)◇	0.530	0.765	—	82.81/85.17	82.53/85.30	Unaligned
MISA (B)*	0.558	0.748	51.45	82.14/85.09	82.44/84.94	Aligned
Self-MM (B)*	0.534	0.764	53.32	84.37/85.28	84.42/85.06	Aligned
MGHF (B)	0.528	0.767	53.70	85.25/85.30	85.09/84.86	Aligned

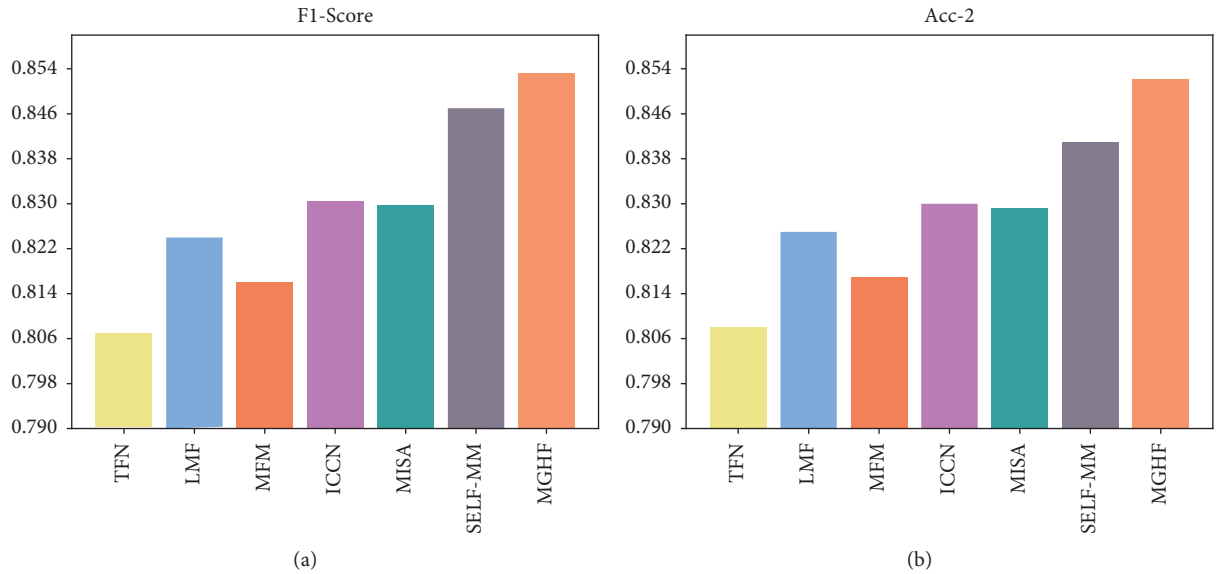

Figure 4: Comparison of the performance of each baseline model. (a) F1-score. (b) Acc-2.

5.2.1. Representational Interaction. First, for cross-modal attention interactions, we conducted the following experiments. The first group was performed for the interaction between two modalities, and we did not consider acoustic-based text features and visual-based text features because this would make the text modality so heavily dominated that

Table 4: Performance tables for different cross-modal notes on MOSI and MOSEI datasets.

Task	MOSI				MOSEI			
	MAE (↓)	Corr (↑)	Acc-2 (↑)	F1-score (↑)	MAE (↓)	Corr (↑)	Acc-2 (↑)	F1-score (↑)
f_{a_v}	1.442	0.210	53.81	46.48/	1.315	0.197	60.13/61.25	60.48/59.38
f_{v_a}	1.321	0.233	62.33	57.85/58.94	1.244	0.182	58.48/59.47	61.63/61.48
f_{a_t}	0.896	0.393	68.52	64.44/65.38	0.815	0.213	64.15/63.18	64.85/64.25
f_{v_t}	0.901	0.384	71.49	69.12/67.20	0.843	0.241	63.48/63.14	63.54/63.89
$f_{v_t} + f_{a_v}$	0.976	0.223	73.62/71.40	71.04/64.31	0.821	0.213	61.84/62.37	61.23/61.66
$f_{a_t} + f_{v_a}$	0.957	0.381	74.22/72.67	71.04/65.86	0.784	0.230	63.24/62.56	61.05/60.72
$f_{a_t} + f_{v_t}$	0.819	0.486	76.80/76.01	75.72/74.84	0.763	0.361	72.18/72.56	74.37/74.03

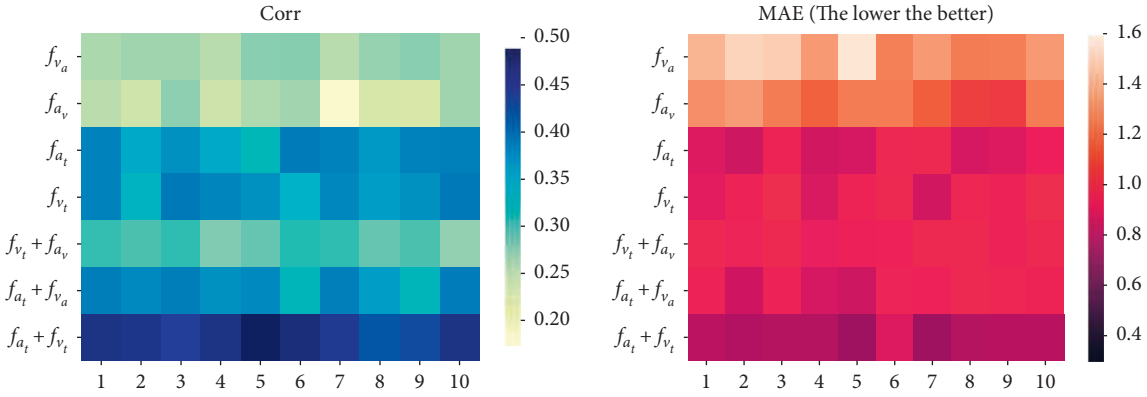

Figure 5: Visualization of different cross-modal interactions and their combined performance. (a) Corr. (b) MAE (the lower the better).

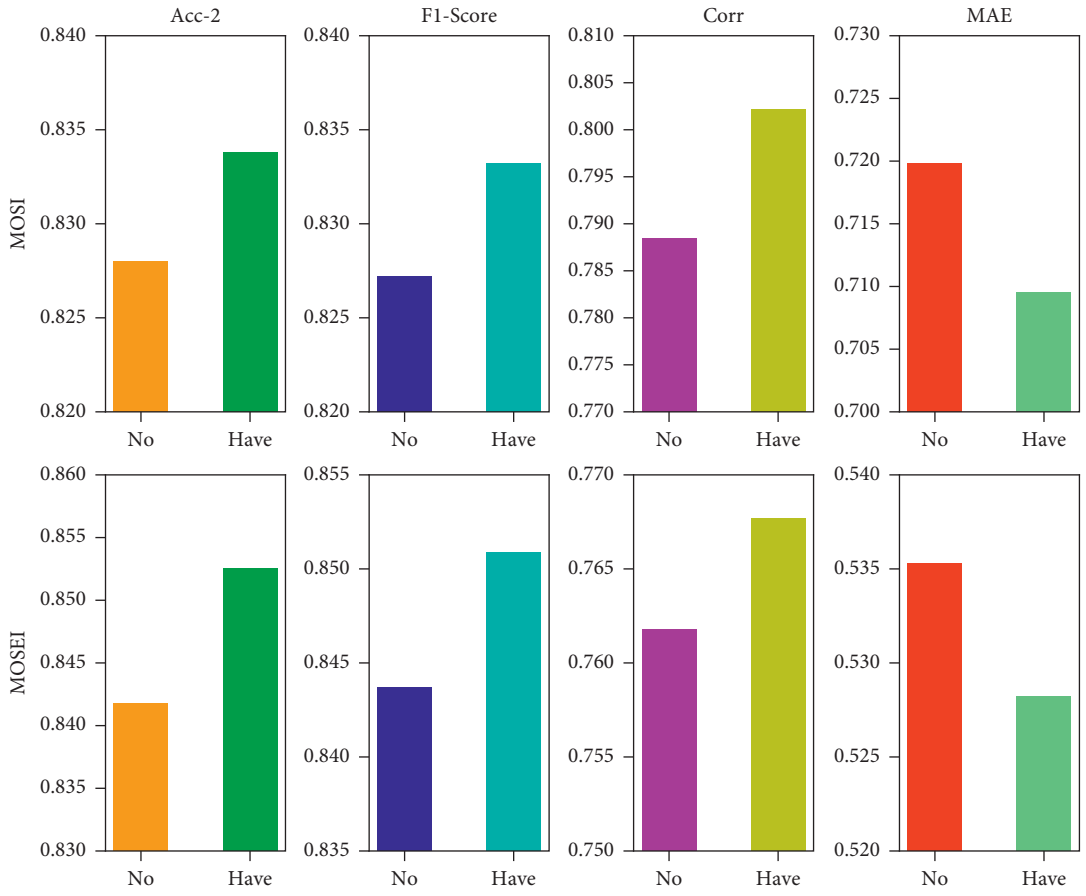

Figure 6: Gated cyclic hierarchical fusion network performance visualization.

Table 5: Ablation study results of fusion strategies on MOSI and MOSEI datasets.

Task	MOSI				MOSEI			
	MAE (↓)	Corr (↑)	Acc-2 (↑)	F1-score (↑)	MAE (↓)	Corr (↑)	Acc-2 (↑)	F1-score (↑)
MGHF_*w/o(pc)*	0.714	0.793	82.88	82.63	0.538	0.758	84.27	84.48
MGHF_*LSTM*	0.712	0.800	83.27	82.94	0.530	0.767	85.22	84.89
MGHF_*original*	0.709	0.802	83.38	83.32	0.528	0.767	85.25	85.09

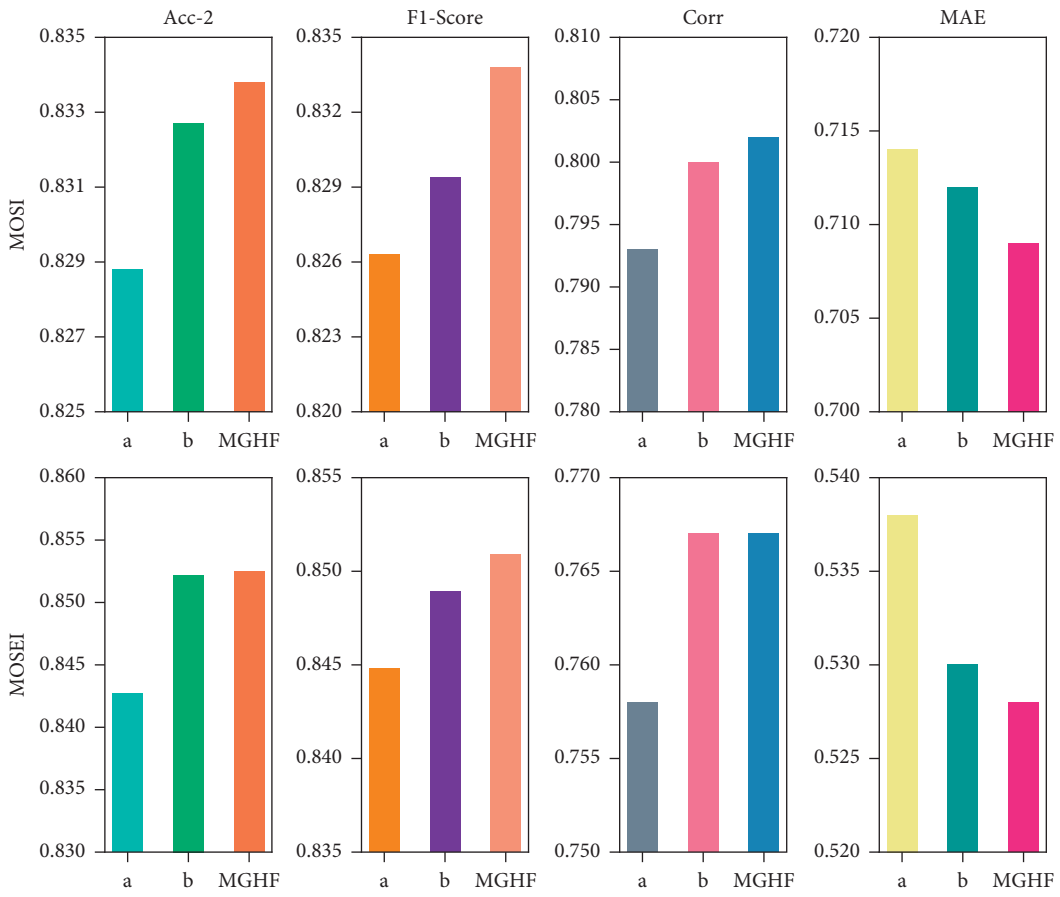

Figure 7: Ablation research in fusion strategies.

modal independence would be reduced or even disappear. The cross-modal attention between nonverbal sequences is hardly satisfactory, probably due to the characteristics of nonverbal sequence data. Acoustic, visual, and textual cross-modal attention seems to play an important role, which is consistent with previous studies [9, 12]. The second set of experiments was conducted after combining the cross-modal interaction representations obtained in the first set, which could help us elucidate whose combination of cross-modal interactions is more beneficial for the MSA task. In Table 4, it seems apparent that the combination of text-based acoustic and text-based visual representations performs the best. We believe this is partly because the text modality enhances the complementary acoustic and visual information, providing additional cues for semantic and affective disambiguation [37], and partly because it preserves the independence of the acoustic and visual modalities. We visualized this part of the experimental index scores for ten randomly selected samples from the MOSI test set (see Figure 5), and similar results were observed on MOSEI.

5.2.2. Gated Recurrent Hierarchical Fusion Network Effectiveness. To verify the reliability of our proposed gated cyclic hierarchical fusion network, we will perform the multimodal sentiment analysis task under the same conditions without this fusion strategy. For visual comparison, two representative metrics from the classification and regression tasks are selected for evaluation, while the evaluation results are visualized. It is worth noting that among these metrics, higher scores imply better performance, except for the MAE metric. The results are shown in (Figure 6). Specifically, the gating mechanism effectively removes the redundant information contained in the previous stage. This not only implies that the representations obtained by the model in the prediction stage are inclusive of the potential

Table 6: Effect of inputs on prediction results for different gated cyclic hierarchical fusion networks on MOSI and MOSEI datasets.

Input	MOSI				MOSEI			
	MAE (↓)	Corr (↑)	Acc-2 (↑)	F1-score (↑)	MAE (↓)	Corr (↑)	Acc-2 (↑)	F1-score (↑)
$f_a + f_{a_t} + f_{v_t}$	0.887	0.641	78.56	78.42	0.699	0.548	80.92	80.33
$f_v + f_{a_t} + f_{v_t}$	0.852	0.728	80.74	80.57	0.644	0.615	82.14	82.06
$f_t + f_{a_t} + f_{v_t}$	0.709	0.802	83.38	83.32	0.528	0.767	85.25	85.09

representations of each modality but also helps us clarify the need for representation interaction learning at a later stage.

In addition, we also conduct ablation experiments of the fusion strategy (as shown in Table 5). In this experiment, we do not combine the resulting text-based visual modality, text-based acoustic modality, and text modality. The settings are marked as "a" in Figure 7 and MGHF_*w/o(pc)* in Table 5. At the same time, we replace Bi-GRU in the fusion network with Bi-LSTM neural network. This setting is marked as "b" in Figure 7 and MGHF_*LSTM* in Table 5. As mentioned before (Section 3.4), only in this experiment does Bi-GRU achieves comparable or even better performance on some metrics.

As shown in the previous section (Section5.2.1) the combined contribution of text-based sound representations f_{a_t} and text-based visual representations f_{v_t} is the highest. We used these two representations combined with the initial representations f_a, f_v, and f_t to evaluate which set performs best for the hierarchical fusion network (see Table 6). It is easy to see that the combination of f_{a_t} and f_{v_t} with the textual representation f_t, which is the input to our gated recurrent hierarchical fusion network, performs best.

6. Conclusions

In this paper, we propose a complete solution for multimodal sentiment analysis, MGHF, which differs in two main parts: modal representation and modal fusion. By using distribution matching in the representation learning phase, the neighbouring modalities are made to contain potential representations of the companion modalities to achieve modal information interaction in time series. Meanwhile, we design a gated recurrent hierarchical fusion network in the fusion phase through the intermodal representation interactions performed in the later fusion phase. It eliminates redundant modal representations and retains those valid for prediction in the final stage, making the prediction results closer to the actual scores. We show that our model is intensely competitive with previous complex baselines through extensive experiments on two publicly available datasets.

Data Availability

The data used to support the findings of this study are available on the following address. Dataset Address: https://immortal.multicomp.cs.cmu.edu/raw_datasets/processed_data/.

Conflicts of Interest

The author(s) declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported by the National Key Research and Development Program of China (Grant No. 2019YFB1404700).

References

[1] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: an adversarial representation learning and graph fusion network for multimodal fusion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 1, pp. 164–172, Washington, DC, USA, April 2020.

[2] W. Yu, X. Hua, F. Meng, and Y. Zhu, "Ch-sims: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[3] D. K. Nugroho, "US presidential election 2020 prediction based on Twitter data using lexicon-based sentiment analysis," in *Proceedings of the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 136–141, IEEE, Noida, India, January 2021.

[4] S. Garg, "Drug recommendation system based on sentiment analysis of drug reviews using machine learning," in *Proceedings of the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 175–181, IEEE, Noida, India, January 2021.

[5] D. S. Chauhan, M. S. Akhtar, A. Ekbal, and P. Bhattacharyya, "Context-aware interactive attention for multi-modal sentiment and emotion analysis," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5647–5657, Hong Kong, China, November 2019.

[6] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: a survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[7] S. Verma, C. Wang, L. Zhu, and W. Liu, "Deepcu: integrating both common and unique latent information for multimodal sentiment analysis," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 3627–3634, Vienna, Austria, August 2019.

[8] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, "Tensor Fusion Network for Multimodal Sentiment Analysis," 2017, https://arxiv.org/abs/1707.07250.

[9] Y. H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, and R. Salakhutdinov, "Multimodal transformer for

unaligned multimodal language sequences," *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019, p. 6558, 2019.

[10] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: modality-invariant and-specific representations for multimodal sentiment analysis," *Proceedings of the 28th ACM International Conference on Multimedia*, New York, NY, USA, 2020.

[11] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, Article ID 10790, 2021, May.

[12] N. Q. Pham, J. Niehues, T. L. Ha, and A. Waibel, "Improving zero-shot translation with language-independent constraints," *WMT*, vol. 13, 2019.

[13] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," *AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 8992–8999, 2020, April.

[14] A. Zadeh, R. Zellers, E. Pincus, and L. P. Morency, "Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[15] A.A. B. Zadeh, P. L. Paul, P. Soujanya, E. Cambria, and M. Louis-Philippe, "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*, vol. 1, Vancouver, Canada, 2018.

[16] S. Schmidhuber and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] E. C. SoujanyaPoria, D. Hazarika, N. Majumder, Amir Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos,"vol. 1, pp. 873–883, in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 1, Association for Computational Linguistics, Vancouver, Canada, 2017.

[18] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, A. N. Vembu, and R. Prasad, "Emotion recognition using acoustic and lexical features," in *Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association*, South Brisbane, QLD, Australia, April 2012.

[19] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L. P. Morency, "Memory fusion network for multi-view sequential learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018, April.

[20] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L. P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," 2018, https://arxiv.org/abs/1806.00064.

[21] W. Dai, S. Cahyawijaya, Y. Bang, and P. Fung, "Weakly-supervised Multi-Task Learning for Multimodal Affect Recognition," 2021, https://arxiv.org/abs/2104.11560.

[22] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: dynamically adjusting word representations using nonverbal behaviors," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 7216–7223, 2019.

[23] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[24] X. Shu, J. Yang, R. Yan, and Y. Song, "Expansion-squeeze-excitation fusion network for elderly activity recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[25] X. Shu, Q. Guo-Jun, J. Tang, and J. Wang, "Weekly-shared deep transfer networks for heterogeneous-domain knowledge propagation," *ACM International Conference on Multimedia (ACM MM)*, Brisbane, Australia, 2015.

[26] "Generalized deep transfer networks for heterogeneous-domain knowledge propagation," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 12, no. 4s, 2016.

[27] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine, "Supervised Multimodal Bitransformers for Classifying Images and Text," 2019, https://arxiv.org/abs/1909.02950.

[28] C. Xi, G. Lu, and J. Yan, "Multimodal sentiment analysis based on multi-head attention mechanism," in *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*, pp. 34–39, Haiphong City, Viet Nam, January 2020.

[29] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018, https://arxiv.org/abs/1810.04805.

[30] R. Aharoni and Y. Goldberg, "Unsupervised domain clusters in pretrained language models," 2020, https://arxiv.org/abs/2004.02105.

[31] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*Long Beach, CA, USA, 2017.

[32] J. Pennington, R. Socher, and C. D. Manning, "Glove:Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, 2014.

[33] G. Degottex, J. Kane, T. Drugman, TuomoRaitio, and S. Scherer, "COVAREP—a collaborative voice analysis repository for speech technologies," in *Proceedings of the 2014 Ieee International Conference on Acoustics, Speech and Signalprocessing (Icassp)*, pp. 960–964, IEEE, Florence, Italy, May 2014.

[34] E. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, Oxford University Press, Oxford, UK, 1997.

[35] Y. H. H. Tsai, P. P. Liang, A. Zadeh, L. P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proceedings of the International Conference on Representation Learning*, New Orleans, LA, USA, February 2019.

[36] W. Rahman, M. K. Hasan, S. Lee et al., "Integrating multimodal information in large pretrained transformers," in *Proceedings of the conference. Association for Computational Linguistics*, vol. 2020, p. 2359p. 2359, July 2020.

[37] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," pp. 689–696, 2011, https://icml.cc/2011/papers/399_icmlpaper.pdf.