scientific reports

OPEN

Multimodal GRU with directed pairwise cross-modal attention for sentiment analysis

Zhenkai Qin[1], Qining Luo[1], Zhidong Zang[2] & Hongpeng Fu[3]✉

Multimodal sentiment analysis combines text, audio, and visual signals to understand human emotions. However, current methods often face challenges in handling asynchronous signals and capturing long-term dependencies between different modalities. Early techniques that merge multiple modalities often introduce unnecessary complexity, while newer methods that treat each modality separately may miss important relationships between the signals. Transformer-based models are effective, but they are typically too resource-heavy for practical use. To overcome these issues, we introduce the multimodal GRU model (MulG), which uses a cross-modal attention mechanism to better synchronize the different signals and capture their dependencies. MulG also employs GRU layers, which are efficient in handling sequential data, making it both accurate and computationally efficient. Extensive experiments on datasets such as CMU-MOSI, CMU-MOSEI, and IEMOCAP demonstrate that MulG outperforms existing methods in accuracy, F1 score, and correlation. Specifically, MulG achieves 82.2% accuracy on CMU-MOSI's 7-class task, 82.1% on CMU-MOSEI, and 90.6% on IEMOCAP's emotion classification. Further ablation studies show that each component of the model contributes significantly to its overall performance. By addressing the limitations of previous approaches, MulG offers a practical and scalable solution for applications like analyzing user-generated content and improving human-computer interaction.

Understanding human emotions is a cornerstone of affective computing, with applications spanning human-computer interaction, social media analytics, and mental health diagnostics. Multimodal sentiment analysis, a discipline that integrates textual, auditory, and visual signals, has emerged as a powerful approach to decoding emotions with greater precision and depth. Using complementary information across modalities, it promises to uncover insights that unimodal analysis often misses. However, this field faces persistent challenges. How can we effectively fuse asynchronous and heterogeneous data streams? How do we capture long-range dependencies between modalities without overwhelming computational resources? These questions underscore the complexity of multimodal sentiment analysis.

Existing methods attempt to address these issues, but remain limited in key aspects. Early fusion approaches combine features at the input stage, resulting in high-dimensional representations that are computationally expensive and challenging to optimize. In contrast, late-fusion methods integrate the outputs of unimodal models, often neglecting critical intermodal interactions. Transformer-based architectures, while excelling in modeling dependencies through self-attention, incur significant computational costs, particularly when applied to large-scale real-world datasets. These limitations hinder their scalability and practicality for real-time applications.

In response, we propose the Multimodal GRU (MulG) model, a novel architecture designed to overcome these challenges by unifying temporal sequence modeling with an innovative cross-modal attention mechanism. At the heart of MulG lies a directed pairwise cross-modal attention mechanism, which dynamically aligns asynchronous features across textual, auditory, and visual streams. Unlike traditional methods, this mechanism excels at identifying interdependencies between modalities by attending to relevant features at different time steps, capturing interactions that would otherwise be lost. Complementing this is the GRU network, which efficiently processes temporal dependencies within each modality, offering a balance between computational efficiency and modeling capacity. Together, these components form a robust feature fusion framework, integrating residual connections to enhance stability and accuracy.

Our contributions are as follows.

[1]College of Information Technology, Guangxi Police College, Juntang Street, Nanning, Guangxi, China. [2]School of Social Development, Yangzhou University, Yangzhou 225009, China. [3]Khoury College of Computer Science, Northeastern University, Seattle, WA, USA. ✉email: fu.hongp@northeastern.edu

- We introduce a directed pairwise cross-modal attention mechanism to dynamically synchronize asynchronous features across modalities, enabling the effective capture of nuanced interdependencies.
- We incorporate GRU layers to efficiently model temporal dependencies within each modality, achieving a balance between performance and computational cost.
- We propose a novel feature fusion framework that combines cross-modal attention and temporal modeling, improving the overall performance and robustness of multimodal sentiment analysis.

We evaluated MulG on three widely used datasets: CMU-MOSI, CMU-MOSEI, and IEMOCAP. Experimental results demonstrate that MulG consistently outperforms state-of-the-art methods in metrics such as accuracy, F1 score, and correlation. Detailed ablation studies validate the indispensability of each component, highlighting the robustness and adaptability of the proposed architecture. By addressing the limitations of existing methods and demonstrating strong performance under diverse conditions, MulG offers a scalable solution for multimodal sentiment analysis in real-world scenarios.

Related work

Multimodal sentiment analysis approaches

Multimodal Sentiment Analysis requires the integration of information between the textual, auditory, and visual modalities. Early fusion techniques, which combine features at the input stage, are intuitive but often result in high-dimensional representations, increasing computational costs and reducing interpretability[1,2]. Late fusion approaches, by integrating unimodal outputs at the decision stage, mitigate dimensionality issues, but tend to overlook intricate intermodal interactions[3,4]. Hybrid strategies, such as Tensor Fusion Networks (TFNs)[5], attempt to balance these extremes by modeling both intra-modal and inter-modal dynamics, yet their reliance on tensor operations introduces scalability concerns. These methods illustrate the ongoing challenge of designing scalable and effective fusion mechanisms for multimodal data.

Transformer and LSTM architectures for multimodal analysis

Transformer-based architectures have gained traction due to their ability to model long-range dependencies through self-attention mechanisms[6,7]. Models like MulT[8]leverage cross-modal attention to align features across modalities, achieving strong performance in benchmark datasets. However, their computational complexity, particularly when dealing with long sequences or large-scale datasets, limits their practical deployment in real-world scenarios. In contrast, LSTM-based models, while efficient for temporal sequence modeling, struggle to capture complex intermodal interactions due to their sequential nature and limited capacity to process asynchronous data streams[9,10]. This dichotomy highlights the need for methods that balance efficiency with the ability to handle multimodal interactions effectively.

Cross-modal attention mechanisms and lightweight models

Recent works have explored cross-modal attention mechanisms to address the challenges of multimodal alignment and interaction. Although traditional self-attention mechanisms focus on global dependencies, pairwise directed attention methods[11]offer a more targeted approach, dynamically aligning asynchronous data streams. Furthermore, lightweight architectures, such as GRU-based networks[12], have emerged as efficient alternatives for tasks that require temporal modeling. However, these methods often fail to capture nuanced interdependencies between modalities or require additional mechanisms to handle noisy or incomplete data.

Contributions of the proposed model

Our work builds on these foundations by introducing a Multimodal GRU (MulG) model that integrates a directed pairwise cross-modal attention mechanism. Unlike Transformer-based approaches, which rely heavily on self-attention, our model adopts a more efficient structure by combining GRU layers with targeted cross-modal attention. This design not only reduces computational costs, but also dynamically aligns asynchronous features, addressing a key limitation of LSTM-based models. Furthermore, the inclusion of residual connections enhances the robustness of the model, enabling it to handle noisy or incomplete modalities, a challenge that is often overlooked in existing studies. By demonstrating strong performance on the CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets, our model highlights the potential for scalable, real-world applications such as user-generated content analysis and human-computer interaction.

Method

Feature preparation and preprocessing

Before performing cross-modal fusion, it is essential to preprocess the raw data from each modality (text, audio, and visual) to extract meaningful features. Each modality requires specific preprocessing steps tailored to the characteristics of the data type, ensuring that the features are properly aligned and ready for further analysis. Below is a detailed description of the preprocessing steps for each modality.

For the text modality, the raw text is first tokenized into individual words or subword units. Each word or token is then mapped to a dense vector representation using pre-trained word embeddings such as word2vec. These embeddings capture semantic relationships between words and are crucial for understanding contextual meaning. Once the tokens are transformed into embeddings, a Convolutional Neural Network (CNN) is applied to the tokenized sequences. The CNN uses convolutional filters of varying sizes to capture local patterns and higher-level n-grams within the text, which may indicate sentiment or specific intent. These filters allow the model to learn important features from sequences of words, such as key phrases or linguistic structures. The resulting feature maps are flattened and passed through fully connected layers to obtain a final feature vector, which represents the entire input text and is ready for further processing in the multimodal framework.

In the case of the audio modality, the raw audio signals are processed using openSMILE, a toolkit specifically designed for extracting various audio features. These include Mel-frequency cepstral coefficients (MFCCs), pitch, and other spectral features, all of which are essential for capturing characteristics like emotional tone, speech patterns, and prosody. These features are extracted from the raw waveform to provide a compact, meaningful representation of the audio. After feature extraction, the data is passed through a fully connected neural network to further refine the representations. The network learns to focus on the most salient features, such as emotional expressions in speech or voice tone, and produces a high-level feature vector that encapsulates the important auditory information for downstream tasks.

For the visual modality, raw video frames or images are processed using a 3D Convolutional Neural Network (3D-CNN). The 3D-CNN is particularly effective for handling spatiotemporal data, such as video, as it applies convolutional operations not only across the spatial dimensions (height and width) but also along the temporal dimension (time). This enables the model to capture both the appearance of objects in each frame and the motion patterns across frames. The 3D-CNN extracts hierarchical features from the video, detecting complex patterns such as moving objects or temporal changes in the scene. The output feature maps are then pooled using max pooling to reduce the dimensionality while preserving important spatial and temporal features. These pooled features are then passed through fully connected layers to generate a final feature vector that represents the visual content, which is ready to be fused with the features from the other modalities.

These preprocessing steps ensure that each modality-text, audio, and visual-has been appropriately transformed into a feature vector that captures the essential information. The feature extraction processes are tailored to each modality's characteristics and are designed to highlight the most relevant patterns for the task at hand. After preprocessing, these feature vectors are ready for integration during the cross-modal fusion stage.

Feature fusion implementation

In the MulG model, feature fusion plays a crucial role in integrating information from multiple modalities to improve the performance of sentiment analysis. The feature fusion process in our model involves several key steps: cross-modal processing, feature fusion, and residual connections and output. Each of these steps is designed to ensure that the information from different modalities is effectively combined and utilised for the final sentiment prediction. The process is illustrated in Figure 1.

Cross-modal processing

After feature extraction, cross-modal processing is performed to capture interactions between the modalities. This is visualised in the second block of Figure 1. In the MulG model, we employ a directed pairwise cross-modal attention mechanism, which allows the model to learn intermodal dependencies by attending to relevant features at different time steps[9,12]. For example, features from the audio modality might attend to text features at certain time steps, capturing important cross-modal relationships. The directed attention mechanism ensures that the dependencies between the modalities are effectively modelled, even when their time steps are asynchronous (for example, audio might have a higher frequency than text)[13].

Cross-modal attention mechanism overview: In a typical multimodal setting, each modality (such as audio, text, or video) has its own time scale, meaning that the sequences generated from each modality are not synchronized. For instance, video frames are usually sampled at a much higher frequency than audio samples or text tokens, creating a misalignment in the temporal sequence of the data. The "cross-modal attention mechanism" is designed to dynamically align these time steps, ensuring that relevant information from each modality is appropriately integrated at each corresponding moment in time. This is crucial for multimodal systems, as it helps in synchronizing and correlating data from different sources that operate asynchronously.

Each pair of modalities is handled by the cross-modal GRU networks, where the attention mechanism directs the GRU to focus on the most relevant cross-modal features[14]. The cross-modal attention mechanism aligns

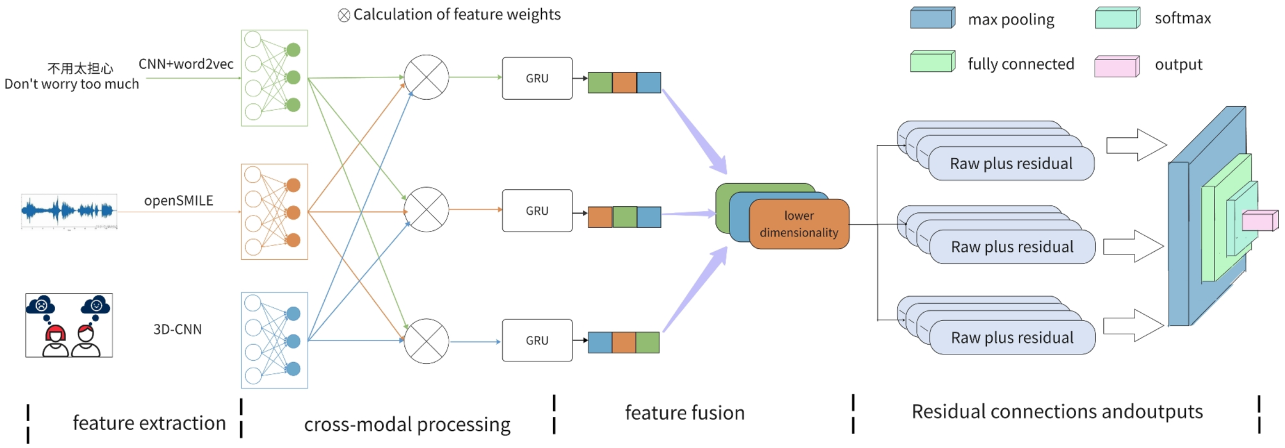

Fig. 1. The architecture of the MulG model consists of four main components: feature extraction, cross-modal processing, feature fusion, and residual connections. The model processes input sequences from text, audio, and image modalities to perform sentiment prediction.

the time steps of the different modalities, and the GRU network processes the attended features to capture both intramodal and intermodal dependencies. Furthermore, the update process of the GRU network, which is used to capture the temporal dependencies of the model, is shown in Figure 2. The GRU uses two gates, the reset gate and the update gate, to manage the flow of information through the network. This structure allows the GRU to retain important temporal information and manage long-term dependencies in sequential data.

As illustrated in Figure 2, the update process of the GRU is described by the following equations:

$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1}) \tag{1}$$

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1}) \tag{2}$$

$$\tilde{h}_t = \tanh(W_{xh}x_t + W_{hh}(r_t \circ h_{t-1})) \tag{3}$$

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \tilde{h}_t \tag{4}$$

r_t is the reset gate, controlling the influence of the previous hidden state. It decides how much of the previous hidden state should be forgotten. A value close to 0 means that the network “forgets” the previous state, and focuses more on the current input. z_t is the update gate, determining how much of the candidate hidden state should be retained. A value close to 1 means that the network keeps the previous hidden state, while a value close to 0 means that it replaces it with the new candidate hidden state. $\tilde{h}_t$ is the candidate hidden state, which is computed using the current input and the previous hidden state, after being transformed by the reset gate. The candidate hidden state is typically passed through a nonlinear activation function, such as $\tanh$, to introduce nonlinearity. h_t is the final hidden state after the update gate decides how much of the candidate hidden state to retain and how much of the previous hidden state to keep.

This structure enables the GRU to effectively manage the flow of information through time, retain relevant temporal information, and handle long-term dependencies in sequential data. The attention mechanism further enhances this process by aligning the time steps of different modalities, ensuring that the GRU processes the most relevant cross-modal features.

In Figure 3, we see that each modality (Language, Video, Audio) is first processed through a Conv1D layer to extract sequential features. The Cross-Modal Transformer (CT) then computes attention from one modality to another. For example, the language (L) modality is in agreement with the video (V) modality and vice versa. This attention mechanism allows each modality to capture relevant information from the other modalities at different time steps. The processed features are then passed into GRU layers to further capture the temporal dependencies within each modality and between different modalities[15].

Paired attention mechanism. For each pair of modalities, the attention mechanism calculates the correlation of each time step in one modality with the other time steps in the modality. This attention score directs the GRU network to focus on the most informative features in each modality[9].

Attention scoring for cross-modal synchronization: Different modalities, such as video, audio, and text, typically operate at distinct sampling rates, leading to sequences of varying lengths (for instance, video frames

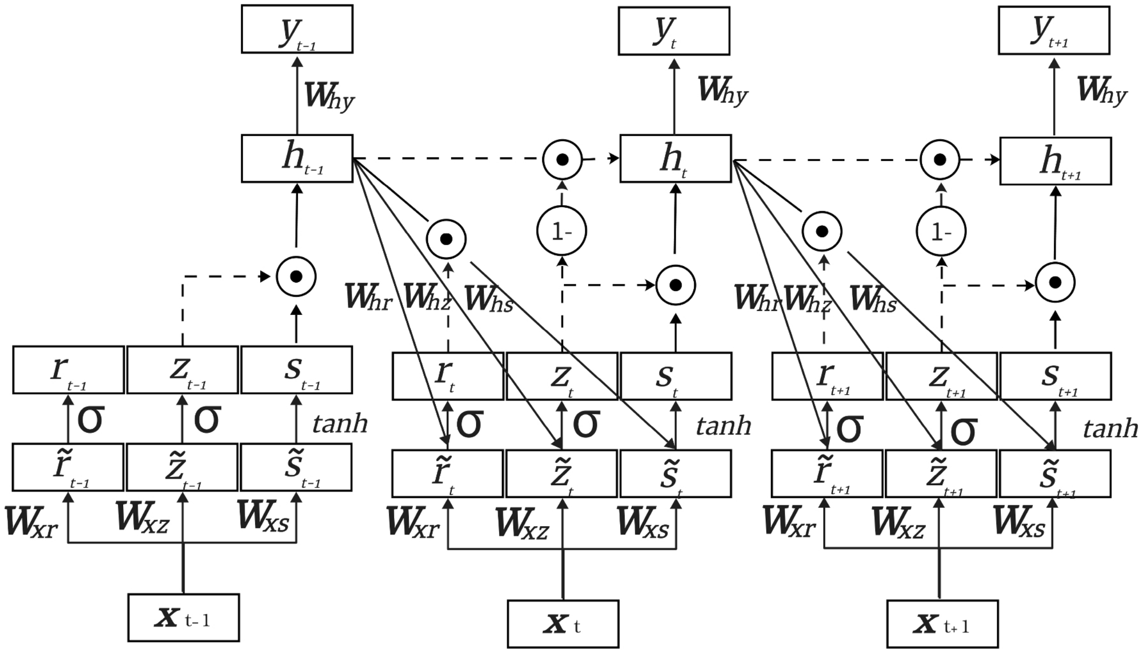

Fig. 2. Detailed GRU network update process showing how the reset gate, update gate, and candidate hidden state interact.

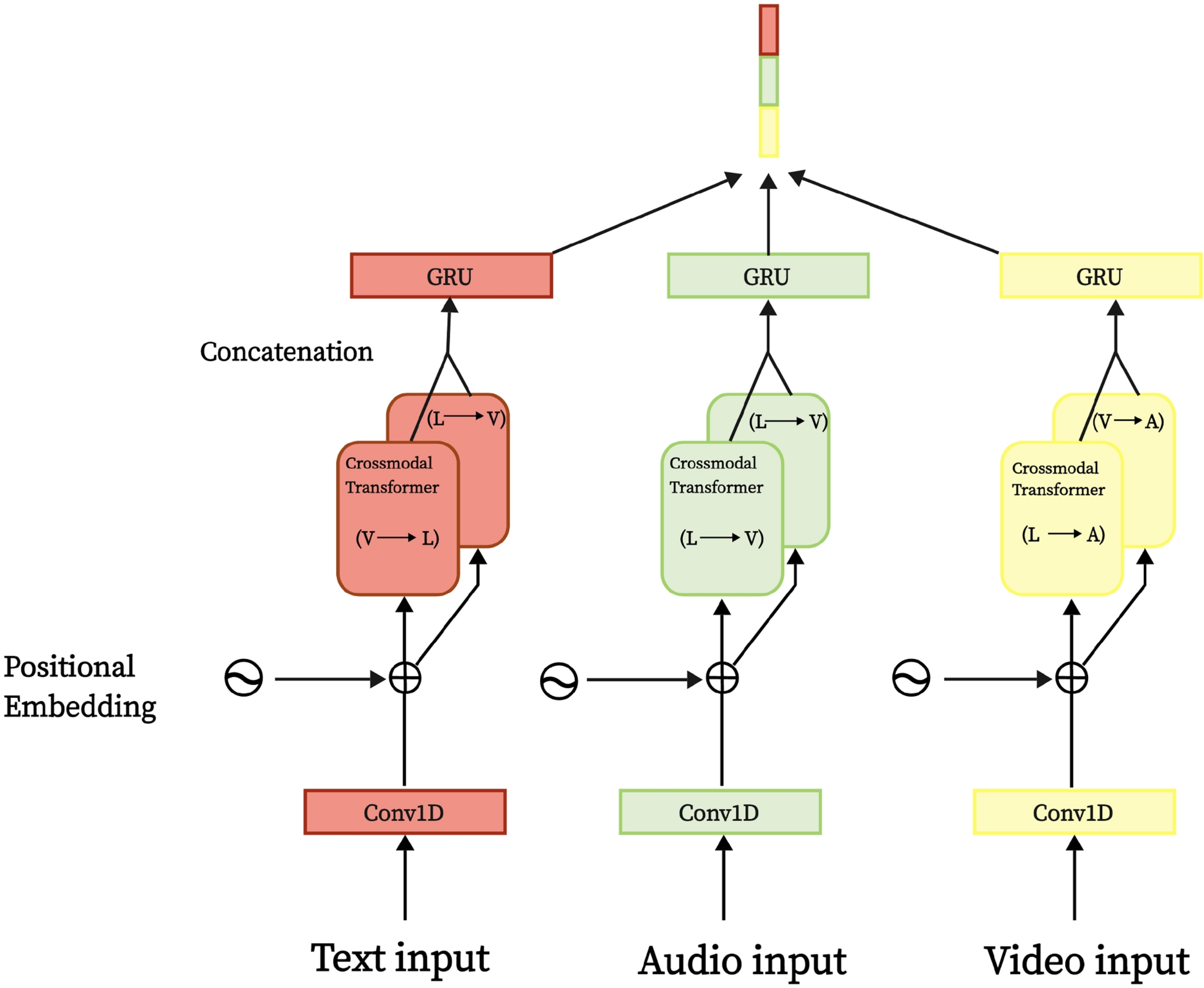

Fig. 3. Cross-modal GRU network for multimodal fusion, showing the cross-modal attention mechanism between language, video, and audio. Each pair of modalities is processed with a Crossmodal Transformer (CT) before feeding into the GRU layers.

tend to occur much more frequently than words or audio samples)[16]. To effectively compute cross-modal attention, we employ a weighted attention mechanism that dynamically aligns the time steps of these different modalities. This mechanism accounts for the varying time scales across modalities, ensuring that the most relevant features are attended to at each time step.

For two modalities, X_1 and X_2, their feature representations are denoted as $X_1 \in \mathbb{R}^{T_1 \times d_1}$ and $X_2 \in \mathbb{R}^{T_2 \times d_2}$, where T_1 and T_2 represent the sequence lengths (e.g., video frames, words, or audio samples), and d_1 and d_2 are the respective feature dimensions. The attention score $e_{t,t'}$ measures the correlation between X_1 at time step *t* and X_2 at time step t', and is computed as follows:

$$e_{t,t'} = (X_1^t)^{\mathrm{T}} W (X_2^{t'}) \tag{5}$$

For example, if X_1 corresponds to a video frame and X_2 represents a word from a sentence, this score quantifies the degree of similarity between a specific video frame at time *t* and a particular word at time t', based on the learned weight matrix *W*. This score acts as a measure of how closely related the features are across modalities at each time step.

Here, *W* is a trainable weight matrix, and X_1^t and $X_2^{t'}$ represent the feature vectors of modalities X_1 and X_2 at time steps *t* and t', respectively[17].

Calculating attentional weights: The attentional weights $\alpha_{t,t'}$ are derived using a softmax function, which normalizes the attention scores across all time steps:

$$\alpha_{t,t'} = \frac{\exp(e_{t,t'})}{\sum_{t'} \exp(e_{t,t'})} \tag{6}$$

For illustration, this equation transforms the attention score $e_{t,t'}$ into a probability distribution, ensuring that the weights across all time steps of modality X_2 sum to one for each time step t in X_1. This normalization helps maintain a consistent, interpretable focus on the most relevant parts of modality X_2 for each time step in X_1.

Weighted feature calculation: Using the computed attentional weights, we obtain a weighted average of the features from modality X_2 to form the aligned feature $\hat{X}_1^t$:

$$\hat{X}_1^t = \sum_{t'} \alpha_{t,t'} X_2^{t'} \tag{7}$$

This weighted feature calculation enables the model to selectively focus on the most relevant features of modality X_2 at different time steps in X_1, effectively aligning asynchronous data streams. Such a mechanism addresses the common challenge of misalignment, especially when dealing with modalities of differing temporal frequencies (e.g., high-frequency video and low-frequency text)[18]. The weighted averaging allows the model to adaptively emphasize the most informative parts of modality X_2 based on the context of modality X_1. This cross-modal alignment is crucial for improving the performance of multimodal systems in real-world applications[19].

Visualization of attention weights: To further illustrate how the directed pairwise cross-modal attention mechanism works, Figure 4shows a heat map representing the attention weights between different time steps in the source and target modalities. This heat map helps visualize which time steps in the text mode are more strongly correlated with specific time steps in the audio or video modalities. The heat map highlights how the model dynamically focuses on relevant features of different modalities during cross-modal interactions[20].

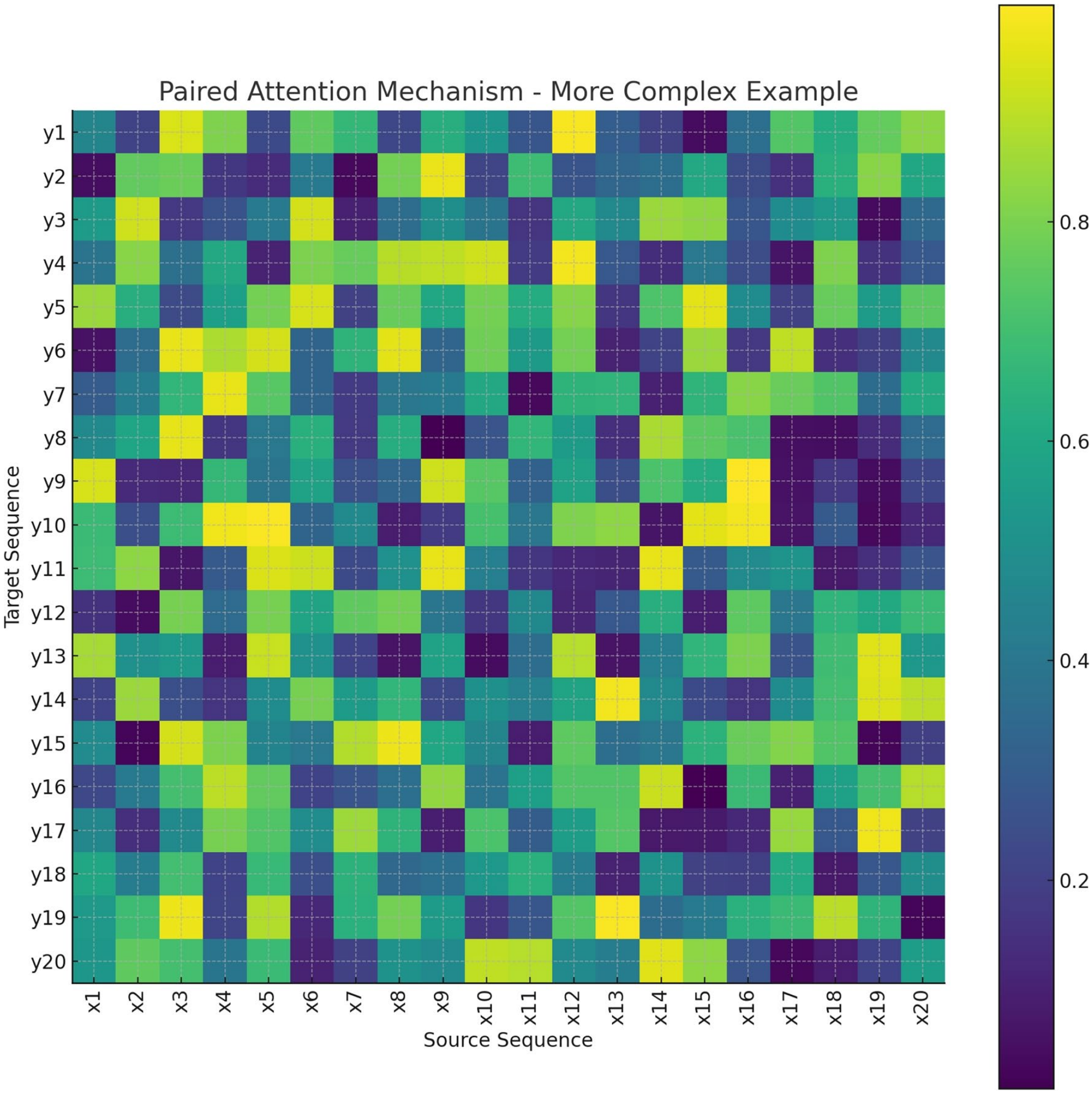

Fig. 4. Heat map showing the attention weights between time steps in the source and target modalities, illustrating how the model focuses on relevant features during cross-modal interactions.

Feature fusion

Once cross-modal processing has been completed, the next step is feature fusion, as shown in the third block of Figure 1. In this stage, the attended and processed features from each modality are concatenated into a single feature vector. This cascading step of the feature ensures that the model captures complementary information from the three modalities: text, audio, and visual.

1. Feature cascading: First, the output of the GRU cross-modal network is cascaded to form a single feature vector. The outputs of the GRU network are h_l (text), h_a (audio), and h_v (visual), and the cascading formula is as follows:

$$h_{\text{concat}} = [h_l; h_a; h_v] \tag{8}$$

where h_{concat} is a spliced feature vector containing combinatorial information from all modal pairs.

2. Dimensionality reduction: Due to the high dimensionality of the spliced feature vector, we apply a fully connected (FC) layer to reduce the size of the feature vector:

$$h_{\text{reduced}} = W_{\text{fusion}} h_{\text{concat}} + b_{\text{fusion}} \tag{9}$$

where W_{fusion} is the weight matrix and b_{fusion} is the bias term. The reduction in dimensionality ensures that the fusion process does not cause an excessive computational burden, and the resultant feature vector is compact and suitable for further processing.

Residual connections

Residual connections are a key component in the MulG model, designed to address the common challenge of gradient vanishing, which can hinder the training of deep neural networks. These connections allow gradients to flow more effectively through the network during backpropagation, leading to faster and more stable convergence. By providing a shortcut for the flow of information between layers, residual connections help preserve important features and prevent the loss of useful information as the network deepens.

In the MulG model, three variants of residual connections are explored, each with distinct configurations that affect how the residuals are integrated into the network. These variants include: (a) post-Layer Normalization (post-LN), (b) pre-Layer Normalization (pre-LN), and (c) ResiDual. Each variant plays a unique role in shaping the learning process and improving the model's performance.

(a) **Post-Layer Normalization (post-LN):** In this variant, the residual connection is applied after the layer normalization step. This means that the output of each layer is first normalized, and then the residual is added to this normalized output. The primary benefit of post-LN is that it helps maintain the scale of the activations after normalization, ensuring that the network can continue learning from the residual information without significant scale distortions. This configuration has been shown to improve training stability, especially for very deep networks.

(b) **Pre-Layer Normalization (pre-LN):** In contrast to post-LN, pre-LN applies the residual connection before the normalization step. In this setup, the input to each layer is normalized first, and then the residual is added. Pre-LN has been found to speed up convergence in some cases, as it ensures that the residual information is injected early into the layer, which can help the network adjust more quickly during training. This approach can sometimes be more effective in scenarios where deeper networks are used, as it provides more direct guidance for the gradient flow.

(c) **ResiDual:** The ResiDual variant represents a hybrid approach that combines both pre-LN and post-LN techniques. This method introduces dual residual paths in the network, where both pre-LN and post-LN residuals are applied, creating a more flexible and robust residual connection. The ResiDual variant aims to leverage the strengths of both approaches, providing more efficient gradient propagation and further enhancing convergence speeds. This variant is particularly useful for complex models that require a fine balance between stability and learning speed, as it can mitigate issues that may arise from relying solely on one normalization method.

Each of these residual variants is depicted in Figure 5, where the different configurations are compared visually. The choice of which residual variant to use depends on the specific characteristics of the model and the task at hand. The effectiveness of these variants in the MulG model highlights their importance in training deep networks and their role in optimizing both convergence speed and model performance.

In summary, residual connections, particularly the use of post-LN, pre-LN, and ResiDual variants, play a critical role in improving the training efficiency and performance of the MulG model. By facilitating better gradient flow and enabling faster convergence, these connections ensure that the model can learn more effectively, even with the added complexity of multimodal feature fusion.

Experiments

Datasets and evaluation metrics

CMU-MOSI and CMU-MOSEI

The CMU-MOSI dataset is designed for human multimodal sentiment analysis and consists of 2,199 short monologue video clips, each containing a single sentence. For both the CMU-MOSI and CMU-MOSEI datasets, each sample is labelled by a human annotator with sentiment scores ranging from −3(strongly negative) to 3 (strongly positive)[21]. We evaluated model performance using various metrics consistent with previous studies: 7-class precision (Acc7: sentiment scores categorised at $\mathbb{Z} \cap [-3, 3]$), dichotomous precision (Acc2: positive / negative sentiments), F1 scores, Mean Absolute Error (MAE) of the scores, and the correlation of the model

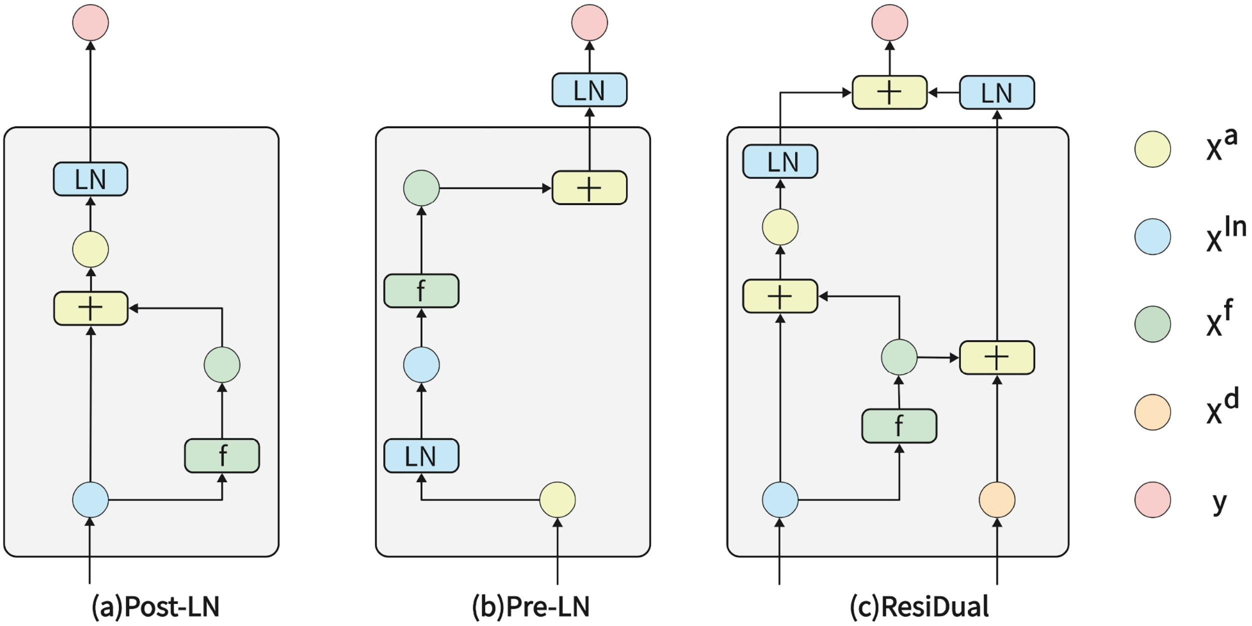

Fig. 5. (**a**) Post-LN: Layer Normalization is performed after Residual connection. (**b**) Pre-LN: Layer Normalization is performed before Residual connection. (**c**) ResiDual: More variants are introduced on the Residual connection by adding additional Layer Normalization layers. All three configurations show the flow of inputs (represented by different colored circles) after passing through the function and Layer Normalization (blue rectangular blocks).

predictions with human annotations. Both tasks are commonly used to benchmark the ability of models to fuse multimodal sentiment information. In addition, we recorded the cross-entropy loss function during training.

IEMOCAP

The IEMOCAP dataset consists of 10,000 videos for human emotion analysis. Four emotions (happy, sad, angry, and neutral) were selected for emotion recognition as suggested by[22]. Unlike CMU-MOSI and CMU-MOSEI, this is a multi-label task, where a person can feel sad and angry at the same time. Its multimodal stream uses a fixed sampling rate of 12.5 Hz for audio signals and 15 Hz for visual signals. We follow the studies of[23] in reporting dichotomous classification accuracy and predicted F1 scores.

Baselines

We chose Early Fusion LSTM (EF-LSTM), Late Fusion LSTM (LF-LSTM), Recurrent Participating Variant Embedding Network (RAVEN) and Multimodal Cyclic Translation Network (MCTN) as the baseline models. The EF-LSTM model fuses the multimodal data in the input stage and then processes them using an LSTM network, which is suitable for scenarios where the intermodal relationships are relatively close[24]. The LF-LSTM model processes the data of each modality separately and extracts high-level features before fusion, which is suitable for scenarios where the intermodal relationships are relatively independent[25]. The RAVEN model fuses multimodal data through a recurrent neural network and variant embedding technique, effectively capturing complex dependencies in the sequence data and improving the accuracy of emotion recognition. The MCTN model utilises a recurrent neural network and a translation mechanism to convert and integrate information between different modalities, establishing deep-level associations between different modalities, and improving the ability of collaborative understanding of multimodal data. To ensure a fair comparison between individual models, all models have roughly the same parameter scales, which not only makes the comparison results fairer but also helps verify the performance of different model architectures under similar conditions. The specific hyperparameter settings and experimental details are shown in the table 1.

Results

Tables 2, table 3, and table 4 present the results of the multimodal sentiment analysis on the CMU-MOSI, MOSEl, and IEMOCAP datasets, respectively, comparing various models in different emotional categories using accuracy (Acc^h), F1-score ($F1^h$), and correlation ($Corr^h$). Our proposed multimodal GRU (MulG) model demonstrates superior performance across all data sets. Specifically, MulG achieves the highest accuracy and F1 scores in multiple categories: The accuracies achieved are Acc_7^h at 40.5% and Acc_2^h at 82. 2% in CMU-MOSI. The accuracies achieved are Acc_7^h at 51.4% and Acc_2^h at 82. 1% in CMU-MOSEI and remarkable results in IEMOCAP with the highest scores in most emotional categories, including happy (90. 6% accuracy) and angry (87. 0% accuracy).

In addition, the MulG model demonstrates a rapid decrease in cross-entropy loss while maintaining high correlation values, further demonstrating its robustness and accuracy in capturing multimodal interactions.

Condition	CMU-MOSEI	CMU-MOSI	IEMOCAP
Epochs	50	100	50
Batch Size	16	128	32
Initial Learning Rate	1e-3	1e-3	1e-3
Optimizer	Adam	Adam	Adam
Transformers Hidden Unit Size	40	40	40
Textual Embedding Dropout	0.3	0.2	0.3
Crossmodal Attention Block Dropout	0.1	0.1	0.2
Output Dropout	0.1	0.1	0.1

Table 1. Hyperparameter Settings for Different Datasets.

Model	(Using cross-modal) CMU-MOSI				(Not using cross-modal) CMU-MOSI			
	Acc_2^h	Acc_7^h	F_1^h	$Corr^h$	Acc_2^h	Acc_7^h	F_1^h	$Corr^h$
EF-LSTM[30]	74.2 ± 1.2	33.4 ± 0.9	1.014 ± 0.024	0.601 ± 0.014	66.2 ± 1.1	29.1 ± 0.8	1.109 ± 0.018	0.586 ± 0.013
LF-LSTM[25]	78.5 ± 1.1	34.2 ± 0.8	1.021 ± 0.019	0.621 ± 0.012	71.7 ± 0.9	29.8 ± 0.7	1.121 ± 0.017	0.598 ± 0.011
RMFN[31]	76.4 ± 1.0	35.2 ± 0.7	0.912 ± 0.030	0.693 ± 0.015	69.7 ± 0.8	30.1 ± 0.6	1.018 ± 0.025	0.649 ± 0.014
MFM[31]	78.9 ± 0.9	36.4 ± 0.7	0.900 ± 0.027	0.614 ± 0.013	72.1 ± 0.8	32.4 ± 0.6	1.052 ± 0.022	0.603 ± 0.012
MulG (our)	82.2 ± 1.0	40.5 ± 0.8	0.821 ± 0.019	0.695 ± 0.014	78.3 ± 0.9	35.6 ± 0.7	0.954 ± 0.020	0.651 ± 0.013

Table 2. Experiments for each model with and without cross-modal use (CMU-MOSI) with confidence intervals.

Model	Acc_2^h	Acc_7^h	F_1^h	$Corr^h$	Acc_2^h	Acc_7^h	F_1^h	$Corr^h$
EF-LSTM[30]	77.6 ± 1.1	46.7 ± 1.2	77.8 ± 1.3	0.614 ± 0.014	70.1 ± 1.0	42.9 ± 1.1	70.4 ± 1.2	0.569 ± 0.015
LF-LSTM[25]	79.8 ± 1.0	47.3 ± 1.0	80.1 ± 1.2	0.653 ± 0.012	71.5 ± 1.0	41.5 ± 1.0	73.2 ± 1.1	0.601 ± 0.014
RMFN[31]	80.1 ± 0.9	45.2 ± 0.8	77.0 ± 1.1	0.564 ± 0.016	72.6 ± 1.0	40.1 ± 0.9	70.6 ± 1.0	0.521 ± 0.017
MFM[31]	76.1 ± 1.0	50.2 ± 1.2	79.4 ± 1.3	0.663 ± 0.014	73.6 ± 0.9	43.9 ± 1.0	71.6 ± 1.2	0.661 ± 0.013
MulG (our)	82.1 ± 1.0	51.4 ± 0.9	82.4 ± 1.0	0.704 ± 0.012	75.6 ± 1.0	44.4 ± 1.1	76.6 ± 1.0	0.676 ± 0.013

Table 3. Experiments for each model with and without cross-modal use (CMU-MOSEI) with confidence intervals.

Model	(Using cross-modal) CMU-MOSEI				(Not using cross-modal) CMU-MOSEI			
	Happy	Sad	Angry	Neutral	Happy	Sad	Angry	Neutral
EF-LSTM[30]	85.5 ± 1.0	80.1 ± 1.1	85.3 ± 0.9	67.9 ± 1.3	80.6 ± 0.9	76.9 ± 1.0	80.6 ± 0.8	60.8 ± 1.4
LF-LSTM[25]	85.9 ± 1.1	76.9 ± 1.0	84.6 ± 1.0	61.7 ± 1.2	80.9 ± 0.9	74.6 ± 1.1	80.1 ± 0.9	57.3 ± 1.3
RMFN[31]	90.1 ± 0.9	82.4 ± 1.0	**87.6 ± 0.8**	67.2 ± 1.3	87.6 ± 0.8	79.8 ± 1.0	**85.6 ± 0.7**	68.1 ± 1.2
MFM[31]	88.1 ± 1.0	85.4 ± 1.1	87.5 ± 1.0	72.1 ± 1.2	82.3 ± 1.0	81.6 ± 0.9	84.5 ± 0.8	68.6 ± 1.1
MulG (our)	**90.6 ± 0.8**	**86.2 ± 0.9**	87.0 ± 1.0	**72.5 ± 1.2**	**88.6 ± 0.7**	**82.3 ± 0.8**	84.9 ± 0.9	**70.6 ± 1.1**

Table 4. Accuracy of individual sentiment recognition under the IEMOCAP dataset (with or without cross-modality) with confidence intervals.

The modal attention mechanism effectively captures long-term dependencies and integrates information across modalities. These results highlight the potential of MulG as a robust baseline model in multimodal sentiment analysis. Furthermore, the MulG model outperforms other models in terms of loss reduction, as shown in Figure 6, where it achieves a faster and more stable decrease in cross-entropy loss compared to other competing models.

Ablation studies

We evaluated the importance of each component of the model through a series of ablation experiments on the CMU-MOSEI and CMU-MOSI datasets. First, we separately removed the textual, audio, and visual modalities from the model to observe their effects on performance. The results indicate that removing any modality significantly reduces the model's accuracy and F1 score, demonstrating that each modality is crucial for enhancing model performance. In the CMU-MOSEI dataset, the binary classification accuracy of the model

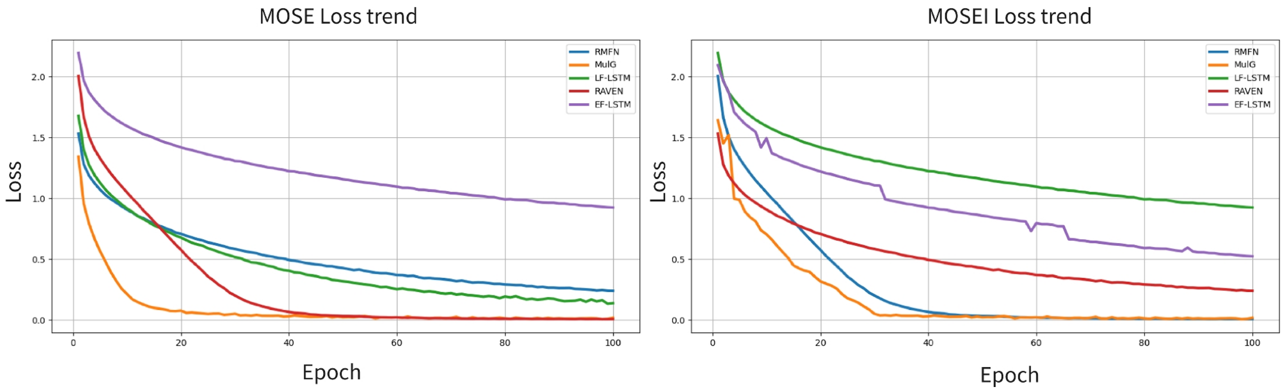

Fig. 6. MulG's comparison with their model in terms of optimizing Loss.

CMU-MOSEI							
	Acc_2^h	Acc_7^h	F_1^h	$Corr^h$	Precision	Recall	MAE
Full Model	82.1 ± 1.0	51.4 ± 1.1	82.4 ± 1.0	0.704 ± 0.012	81.5 ± 1.0	80.8 ± 1.1	0.423 ± 0.015
Text Only Modal	80.2 ± 0.9	50.1 ± 1.0	80.1 ± 1.0	0.673 ± 0.014	79.3 ± 0.9	79.0 ± 1.0	0.435 ± 0.017
Audio Only Modal	79.3 ± 0.8	48.6 ± 0.9	78.9 ± 1.1	0.681 ± 0.013	78.2 ± 0.8	78.0 ± 0.9	0.442 ± 0.016
Visual Modality Only	80.6 ± 0.9	49.3 ± 0.8	77.5 ± 1.0	0.678 ± 0.012	78.9 ± 0.8	77.0 ± 0.9	0.430 ± 0.014
No Cross-Modal Attention	75.6 ± 1.0	44.4 ± 1.2	76.6 ± 1.1	0.676 ± 0.015	74.8 ± 1.0	75.0 ± 1.1	0.460 ± 0.018

Table 5. Accuracy and other metrics for various ablation experiments on the CMU-MOSEI dataset with confidence intervals.

CMU-MOSI							
	Acc_2^h	Acc_7^h	F_1^h	$Corr^h$	Precision	Recall	MAE
Full Model	85.0 ± 1.1	55.0 ± 1.0	85.2 ± 1.0	0.720 ± 0.012	83.0 ± 1.0	82.5 ± 1.0	0.400 ± 0.014
Text Only Modal	82.0 ± 1.0	52.0 ± 1.0	82.0 ± 1.0	0.690 ± 0.013	80.0 ± 0.8	79.5 ± 0.9	0.420 ± 0.015
Audio Only Modal	80.0 ± 0.9	50.0 ± 0.9	79.0 ± 1.1	0.670 ± 0.012	78.5 ± 0.7	78.0 ± 0.8	0.440 ± 0.016
Visual Modality Only	81.0 ± 0.9	51.0 ± 0.8	78.5 ± 1.0	0.680 ± 0.011	79.0 ± 0.8	77.5 ± 0.9	0.430 ± 0.014
No Cross-Modal Attention	76.0 ± 0.9	46.0 ± 1.1	77.0 ± 1.2	0.660 ± 0.014	75.0 ± 0.9	74.5 ± 1.0	0.450 ± 0.016

Table 6. Accuracy and other metrics for various ablation experiments on the CMU-MOSI dataset with confidence intervals.

drops to 80.2% when only the text modality is used, 79.3% when only the audio modality is used, and 80.6% when only the visual modality is used. The complete model achieves a binary classification accuracy of 82.1%. These results are detailed in Table5.

Similarly, on the CMU-MOSI dataset, the model's binary classification accuracy drops to 82.0% when only the text modality is used, 80.0% when only the audio modality is used, and 81.0% when only the visual modality is used. The complete model achieves a binary classification accuracy of 85.0%. These results are shown in Table 6.

To further explore the role of cross-modal attention mechanisms, we conducted experiments without cross-modal attention. In these experiments, we removed the cross-modal attention module and relied only on the unimodal self-attention mechanism or an alternative GRU module for information fusion. The results show a performance degradation, with the binary classification accuracy dropping to 75.6% on the CMU-MOSEI dataset and 76.0% on the CMU-MOSI dataset. However, the model still achieves reasonably good results. This suggests that cross-modal attention is vital for effectively integrating multimodal information, while the unimodal self-attention mechanism or GRU module also has some capability in processing respective modal information.

Finally, we evaluated the effectiveness of the self-attention mechanism compared to traditional RNN modules such as GRU. Replacement of the self-attention mechanism with GRU resulted in a slight decrease in performance, with a binary classification accuracy of 79.3% on the CMU-MOSEI dataset and 80.0% on the CMU-MOSI dataset. These results indicate that the self-attention mechanism has advantages in capturing long-distance dependencies and handling complex multimodal interactions, while the GRU module, although slightly

Hyperparameter	Value
Original Text Dimension (orig_d_l)	50
Original Audio Dimension (orig_d_a)	50
Original Visual Dimension (orig_d_v)	50
Model Output Dimension (output_dim)	10
Number of Attention Heads (num_heads)	4
Layers (layers)	2
Attention Dropout (attn_dropout)	0.1
Audio Attention Dropout (attn_dropout_a)	0.1
Visual Attention Dropout (attn_dropout_v)	0.1
ReLU Dropout (relu_dropout)	0.1
Residual Dropout (res_dropout)	0.1
Output Dropout (out_dropout)	0.1
Embedding Dropout (embed_dropout)	0.1

Table 7. Hyperparameterization of ablation experiments.

inferior, can still handle unimodal information to some extent. The hyperparameters of all components of the ablation experiment are shown in Table7.

In general, these experiments validate the critical role of multimodal information fusion and cross-modal attention mechanisms in enhancing the performance of sentiment analysis models. These experiments not only help us understand the model's working mechanisms, but also provide valuable insights for further model optimization.

Discussion

Model effectiveness and comparisons

The proposed multimodal GRU (MulG) model, which incorporates directed pairwise cross-modal attention, presents a significant advancement in multimodal sentiment analysis. MulG's ability to focus on the interactions between low-level features from different modalities, text, audio, and visual, enables it to capture long-range dependencies across these modalities more effectively than existing models. This is evident from its superior performance in precision, F1 scores, and correlation metrics across the *CMU-MOSI*, *CMU-MOSEI*, and *IEMOCAP* datasets.

MulG's performance stands out in comparison with the Early Fusion LSTM (EF-LSTM) and Late Fusion LSTM (LF-LSTM) models. EF-LSTM struggles to model complex inter-modal relationships, as it fuses modalities at the input stage, failing to capture intricate dependencies. LF-LSTM, on the other hand, processes each modality independently before fusion, which can miss critical interactions. MulG, by leveraging directed cross-modal attention, ensures that the most informative features from each modality are selectively attended to, leading to more effective fusion and improved overall performance.

Furthermore, while the *multimodal Transformer (MulT)* model, proposed by Tsai et al.[33], has demonstrated the power of attention mechanisms for multimodal fusion, it often faces significant computational challenges due to the heavy reliance on the Transformer architecture. The self-attention mechanism in MulT, though powerful, can become inefficient when processing long sequences or large datasets. In contrast, MulG integrates GRU layers, effectively managing temporal dependencies while using cross-modal attention for fusion. This design strikes a balance between high performance and computational efficiency, making MulG a more feasible solution for real-world applications with resource constraints.

Model and parameter adjustments

The effectiveness of MulG is closely related to the careful selection of architectural components and hyperparameters. Oriented pairwise cross-modal attention plays a crucial role in enhancing the ability of the model to capture subtle interdependencies between modalities, which directly contributes to improved sentiment recognition. Ablation studies conducted earlier demonstrated that removing any single modality from the model leads to a significant drop in performance, underlining the importance of multimodal data for accurate sentiment analysis.

Key parameters, such as dropout rates and learning rates, were optimised to ensure the model's robustness and prevent overfitting. The dropout rate in the attention mechanism was set at 0.1, providing a good balance between regularisation and learning capacity[34]. The learning rate was fine-tuned through a grid search and set to 0.001, ensuring gradual convergence without the risk of overshooting[35]. Furthermore, the choice of using two GRU layers with a hidden size of 40 was intended to capture temporal dependencies while maintaining computational efficiency, which is consistent with the findings of previous studies on RNN depth and gradient stability[36]. Batch normalisation and ReLU activation functions were used to stabilise the training process and improve convergence[37].

These adjustments resulted in the strong performance of MulG on datasets such as CMU-MOSI and CMU-MOSEI, validating the impact of careful parameter selection. The success of MulG in these settings also emphasises the critical role of hyperparameter optimisation in multimodal models[38].

Computational efficiency analysis

One of the key challenges of the proposed MulG model is the computational overhead introduced by the cross-modal attention mechanism. Unlike baseline models such as LSTMs or standard RNNs, which process each modality independently, the intermodal attention mechanism computes attention scores between modalities at each time step. This results in a higher computational complexity, as it involves additional matrix multiplications and nonlinear transformations. Specifically, the attention mechanism requires pairwise computations between all modalities at each time step, leading to a quadratic increase in computational demands as the number of modalities and time steps grows.

Although this added complexity might seem prohibitive for large-scale tasks, MulG addresses this concern by incorporating GRU layers to capture temporal dependencies. The GRU layers help reduce the number of parameters needed to model these dependencies compared to fully Transformer-based models, which typically require a larger number of parameters and more computational resources. The use of GRU layers strikes a balance between performance and efficiency, allowing MulG to maintain good performance while being more computationally efficient than Transformer-based approaches.

However, despite the efficiency improvements provided by GRU, the computational cost of the cross-modal attention mechanism remains significant, especially when working with real-time applications or in resource-constrained environments such as mobile or embedded systems. The computational load of the attention mechanism increases with the number of modalities and time steps, which can present challenges in applications requiring low-latency processing.

Limitations and future work

Despite its strong performance, MulG does have limitations that need to be addressed for broader real-world applicability. The model's dependence on high-quality multimodal data makes it susceptible to performance degradation when faced with incomplete or noisy inputs. Real-world data, especially from sources like low-quality recordings or occluded visuals, could result in reduced accuracy. Although MulG performs well on structured and controlled datasets such as CMU-MOSI and CMU-MOSEI, it may struggle when dealing with the variability and noise often encountered in real-world scenarios.

Furthermore, the datasets used for training and evaluation, including CMU-MOSI, CMU-MOSEI, and IEMOCAP, while well-curated, may not fully represent the diversity and complexity of real-world sentiment data. These datasets are often limited in terms of cultural, linguistic, and situational variability, which can affect the model's ability to generalize to diverse real-world contexts. To address this, we plan to incorporate more diverse datasets in future research, which may include data from various cultural backgrounds and real-world social media platforms to enhance the model's generalization ability.

Future research should explore ways to improve MulG's robustness in handling incomplete, noisy, or corrupted data. Investigating the model's ability to perform well with missing or degraded modalities is crucial, as real-world applications often involve such scenarios. Enhancing the model's ability to adapt to these challenges will greatly improve its utility in practical applications. We also plan to explore the use of data augmentation techniques and domain adaptation methods to mitigate the impact of missing or noisy data, enabling the model to perform better under these conditions.

Additionally, although MulG shows impressive performance on structured datasets, there is a need to assess its generalisation capabilities on more diverse and unstructured data sources, such as social media posts or spontaneous dialogues. These data types introduce challenges such as noisy or unaligned inputs, which are common in real-world multimodal systems. Further evaluations of such datasets will provide valuable information on the real-world applicability of the model. We aim to incorporate large-scale unstructured data sources, such as user-generated content on social media, to evaluate the model's ability to handle the complexities and variabilities inherent in real-world multimodal data.

Finally, optimizing MulG's computational efficiency is another avenue for future work. Exploring lighter attention mechanisms, such as sparse attention or low-rank approximation, could help reduce computational demands while preserving performance, making the model more suitable for deployment in resource-constrained environments like mobile or embedded systems. We also plan to explore techniques like model pruning and quantization to further reduce the computational cost and memory footprint of MulG, ensuring its scalability for real-time applications.

In conclusion, future work should focus on improving the robustness of MulG to noisy and incomplete data, evaluating its adaptability to diverse and unstructured datasets, and optimising its computational efficiency for deployment in low-resource settings. These advancements would enhance the practical utility of MulG in a wide range of multimodal sentiment analysis tasks.

Conclusion

This paper introduces the Multimodal GRU (MulG) model for multimodal sentiment analysis. The key innovation of MulG lies in its cross-modal attention mechanism, which allows the model to directly attend to low-level features from different modalities, facilitating more effective integration of multimodal information. By capturing inter-modal stochastic dependencies through a combination of GRU layers and cross-modal attention, MulG offers a robust solution for analysing human multimodal language.

Our experimental results demonstrate that MulG significantly outperforms previous models in sentiment analysis tasks, highlighting its effectiveness in multimodal fusion and temporal dependency capture. These findings position MulG as a promising candidate for practical applications in multimodal sentiment analysis, where high accuracy and computational efficiency are essential.

Moving forward, MulG's ability to handle noisy, incomplete, and unstructured data will be critical for real-world deployment. Future work should focus on improving its robustness to these data challenges and optimising its computational efficiency to ensure its scalability in low-resource environments. By addressing these areas, MulG has the potential to become a leading model for large-scale multimodal real-time sentiment analysis tasks.

Data availability

Our code has been exposed in the repository:https://github.com/bitbitlemon/Multimodal-GRU. The datasets "CMU-MOSI," "CMU-MOSEI," and "IEMOCAP" are all publicly available and widely used in the fields of multimodal sentiment analysis and emotion recognition. Raw data can be obtained from the following address. CMU-MOSI and CMU-MOSEI:https://github.com/CMU-Multicomp-Lab/CMU-MultimodalSDK. IEMOCAP :https://sail.usc.edu/iemocap/.

Received: 6 December 2024; Accepted: 4 March 2025
Published online: 24 March 2025

References

1. Pérez-Rosas, Verónica, Rada Mihalcea, & Louis-Philippe Morency. "Utterance-level multimodal sentiment analysis." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013.
2. Wöllmer, Martin, et al. "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling." (2010).
3. Poria, Soujanya, Erik Cambria, & Alexander Gelbukh. "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis." Proceedings of the 2015 conference on empirical methods in natural language processing. (2015).
4. Zadeh, Amir, et al. "Tensor fusion network for multimodal sentiment analysis." arXiv preprint arXiv:1707.07250 (2017).
5. Liang, Paul Pu, et al. "Multimodal language analysis with recurrent multistage fusion." arXiv preprint arXiv:1808.03920 (2018).
6. Ngiam, Jiquan, et al. "Multimodal deep learning." Proceedings of the 28th international conference on machine learning (ICML-11). (2011).
7. Atrey, Pradeep K., et al. "Multimodal fusion for multimedia analysis: a survey." Multimedia systems 16: 345-379 (2010).
8. Palmer, Martha, Hwa, Rebecca, & Riedel, Sebastian. "Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing." Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. (2017).
9. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
10. Tsai, Yao-Hung Hubert, et al. "Multimodal transformer for unaligned multimodal language sequences." Proceedings of the conference. Association for computational linguistics. Meeting. Vol. 2019. NIH Public Access, (2019).
11. Wu, Dekai, et al. "Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation." Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. (2014).
12. Bahdanau, Dzmitry, Cho, Kyunghyun, Bengio, Yoshua. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).
13. Dey, Rahul, & Salem, Fathi M. "Gate-variants of gated recurrent unit (GRU) neural networks." 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS). IEEE, (2017).
14. Chorowski, Jan K., et al. "Attention-based models for speech recognition." Advances in neural information processing systems 28 (2015).
15. Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. PMLR, (2015).
16. Ba, Jimmy Lei, Kiros, Jamie Ryan, & Hinton, Geoffrey E. "Layer normalization." arXiv preprint arXiv:1607.06450 (2016).
17. Xie, Saining, et al. "Aggregated residual transformations for deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. (2017).
18. Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).
19. Wang, Fei, et al. "Residual attention network for image classification." Proceedings of the IEEE conference on computer vision and pattern recognition. (2017).
20. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. (2016).
21. Zadeh, Amir et al. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* **31**(6), 82–88 (2016).
22. Busso, Carlos et al. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* **42**, 335–359 (2008).
23. Poria, Soujanya, et al. "Context-dependent sentiment analysis in user-generated videos." Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers). (2017).
24. Poria, Soujanya et al. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion* **37**, 98–125 (2017).
25. Baltrušaitis, Tadas, Ahuja, Chaitanya & Morency, Louis-Philippe. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* **41**(2), 423–443 (2018).
26. Busso, C., Deng, Z., & Lee, C. IEMOCAP: Interactive emotional dyadic motion capture database. *Proceedings of the 5th International Conference on Affective Computing and Intelligent Interaction*, 1-6 (2008).
27. Zhou, J., Yu, L. & Lin, Y. Multimodal emotion recognition using deep learning for human-robot interaction. *Journal of Ambient Intelligence and Smart Environments* **10**(5), 557–569. https://doi.org/10.3233/AIS-180523 (2018).
28. Wang, Y., Zhang, L. & Sun, G. RAVEN: Recurrent participating variant embedding network for multimodal emotion recognition. *IEEE Transactions on Affective Computing* **9**(4), 504–515. https://doi.org/10.1109/TAFFC.2017.2731591 (2018).
29. Pham, P., Lee, G. & Yoon, S. MCTN: Multimodal cyclic translation network for multimodal emotion recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision* **1821–1830**, https://doi.org/10.1109/ICCV.2019.00206 (2019).
30. Poria, Soujanya, Cambria, Erik, Hazarika, Devamanyu, et al. "Context-dependent sentiment analysis in user-generated videos." Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 873-883.
31. Li, Xian, Tao, Wei, Wang, Hongwei, et al. "Recurrent Multistage Fusion Network for Multimodal Sentiment Analysis." Proceedings of the 27th ACM International Conference on Multimedia. 2019: 1698-1706.
32. Zadeh, Amir, Lim, Yao-Chong, Liang, Paul Pu, et al. "Memory Fusion Network for Multimodal Sentiment Analysis." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2018: 151-162.
33. Tsai, Yao-Hung Hubert, Liang, Paul Pu, Zadeh, Amir, et al. "Multimodal transformer for unaligned multimodal language sequences." Proceedings of the Association for Computational Linguistics. (2019).
34. Pérez-Rosas, Verónica, Rada, Mihalcea, & Louis-Philippe, Morency. "Utterance-level multimodal sentiment analysis." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). (2013).

35. Poria, Soujanya, Erik, Cambria, & Alexander, Gelbukh. "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis." Proceedings of the 2015 conference on empirical methods in natural language processing. (2015).
36. Ioffe, Sergey, & Christian, Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." Proceedings of the 32nd International Conference on Machine Learning (ICML). (2015).
37. Baltrušaitis, Tadas, Chaitanya, Ahuja, & Louis-Philippe, Morency. "Multimodal machine learning: A survey and taxonomy." IEEE Transactions on Pattern Analysis and Machine Intelligence. (2018).
38. Girdhar, Rohit, Deva, Ramanan. "Attentional pooling for action recognition." Advances in Neural Information Processing Systems. (2017).

Author contributions

Z.Q. and Q.L. were responsible for the conceptualization of the research, Z.Z. conducted the formal analysis, and Z.Q. carried out the investigation. The methodology was developed by Z.Q. and H.F., while H.F. handled the validation. Z.Z. was responsible for the visualization of the figures. The supervision of the project was managed by Q.L. and Z.Z. Z.Q. and Q.L. wrote the original draft, and Z.Q. and H.F. reviewed and edited the manuscript. This format provides a concise yet clear attribution of each author's contributions.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.