



Full length article

Adversarial alignment and graph fusion via information bottleneck for multimodal emotion recognition in conversations

Yuntao Shou^a, Tao Meng^{a,*}, Wei Ai^a, Fuchen Zhang^a, Nan Yin^b, Keqin Li^c

^a College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha, Hunan 410004, China

^b Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi 44737, United Arab Emirates

^c Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

ARTICLE INFO

Keywords:

Adversarial representation learning
Feature fusion
Graph contrastive representation learning
Multimodal emotion recognition in conversations
Information bottleneck

ABSTRACT

With the rapid development of social media and human–computer interaction, multimodal emotion recognition in conversations (MERC) tasks have begun to receive widespread research attention. The MERC task is to extract and fuse complementary semantic information from different modalities to classify the speaker's emotion. However, the existing feature fusion methods usually directly map the features of other modalities into the same feature space for information fusion, which cannot eliminate the heterogeneity between different modalities and make the subsequent emotion class boundary learning more difficult. In addition, existing graph contrastive learning methods obtain consistent feature representations by maximizing mutual information between multiple views, which may lead to overfitting of the model. To tackle the above problem, we propose a novel Adversarial Alignment and Graph Fusion via Information Bottleneck for Multimodal Emotion Recognition in Conversations (AGF-IB) method. Firstly, we input video, audio, and text features into a multi-layer perceptron (MLP) to map them into separate feature spaces. Secondly, we build a generator and a discriminator for the three modal features, respectively, through adversarial representation to achieve information interaction between modalities and eliminate the heterogeneity among modalities. Thirdly, we introduce graph contrastive representation learning to capture intra-modal and inter-modal complementary semantic information and learn intra-class and inter-class boundary information of emotion categories. Furthermore, instead of maximizing the mutual information (MI) between multiple views, we use information bottleneck theory to minimize the MI between views. Specifically, we construct a graph structure for the three modal features respectively and perform contrastive representation learning on nodes with different emotions in the same modality and nodes with the same emotion in different modalities, to improve the feature representation ability of nodes. Finally, we use MLP to complete the emotional classification of the speaker. Extensive experiments show that AGF-IB can improve emotion recognition accuracy on IEMOCAP and MELD datasets. Furthermore, since AGF-IB is a general multimodal fusion and contrastive learning method, it can be applied to other multimodal tasks in a plug-and-play manner, e.g., humor detection.

1. Introduction

The multimodal emotion recognition in conversations (MERC) task is to combine the semantic information of different modal features (e.g., text, video, and audio, etc.) to identify the emotion of the speaker at the current moment [1]. With the continuous development of deep learning technology and computing resources, MERC has also begun to be widely used in many practical social media scenarios. For example, in a human–computer dialogue system, the interactive system can obtain the user's current emotional state according to the data analysis of the human–computer dialogue, and then generate words that fit the

scene. Therefore, accurately identifying the user's current emotional state has high practical application value [2].

However, MERC must eliminate the modality gap of multi-modal heterogeneous data because video, audio, and text feature embeddings in space are inconsistent [3,4]. The current mainstream feature fusion method to eliminate the gap of different modal data is directly mapping them into the same feature space for feature representation [5]. For example, Tensor Fusion Network (TFN) [6] uses the tensor outer product operation to map different modal features into a three-dimensional feature space for the fusion representation of multi-modal feature vectors.

* Corresponding author.

E-mail addresses: shouyuntao@stu.xjtu.edu.cn (Y. Shou), mengtao@hnu.edu.cn (T. Meng), aiwei@hnu.edu.cn (W. Ai), fuchen.zhang@csuft.edu.cn (F. Zhang), nan.yin@mbzuai.ac.ae (N. Yin), lik@newpaltz.edu (K. Li).

<https://doi.org/10.1016/j.infus.2024.102590>

Received 25 October 2023; Received in revised form 11 June 2024; Accepted 16 July 2024

Available online 20 July 2024

1566-2535/© 2024 Elsevier B.V. All rights reserved, including those for text and data mining, AI training, and similar technologies.

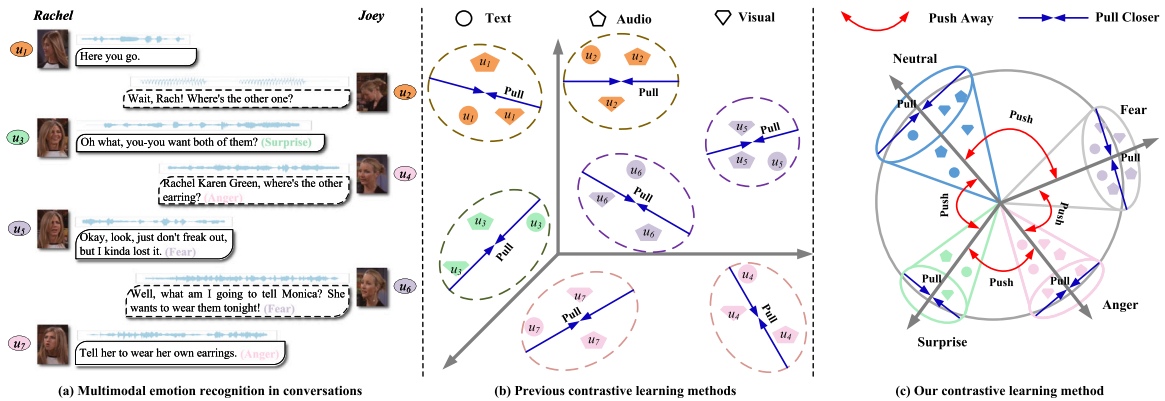


Fig. 1. Illustrative example of the effect of different contrastive learning methods on the multi-modal feature embeddings. Previous works mainly focus on intra-sample contrastive learning, while ignoring inter-sample and inter-class contrastive learning. In our work, we consider both inter-sample and inter-class contrastive learning.

Low-rank Fusion Network (LFN) [7] utilizes low-rank decomposition operations to combine highly correlated feature vectors and fuse three modal features. However, the above methods forcibly map different modal features into a common representation space, which cannot eliminate their heterogeneity. We argue that a suitable feature fusion method should first perform modal alignment and then perform modal fusion.

Another problem with existing deep learning methods is that they fail to capture inter-sample and inter-class semantic information that is differentiated. Taking Fig. 1 as an example, previous contrastive learning work focused on learning interactions and aligning modalities within samples. For instance, Hu et al. [8] proposed Multi-modal Fusion via Deep Graph Convolution Network (MMGCN) to fuse dialogue relations and complementary semantic information of different modalities. Liu et al. [9] proposed a Multi-modal Fusion Network (MFN), which uses an attention mechanism to consider the importance of different modalities and obtains a multi-modal fusion vector with modal interactions. It is difficult for the above methods to learn clear class boundaries between different emotion categories. However, an increasing number of studies have shown that capturing the relationship between samples and emotion categories contributes to better emotion classification. Therefore, as shown in Fig. 1 (c), we construct an inter-modal and inter-class contrastive learning paradigm to learn more discriminative emotional feature representations.

Hence, how to eliminate the heterogeneity between different modalities and capture the intra-modal and inter-modal complementary semantic information and intra-class and inter-class differences is still a problem to be solved.

Furthermore, existing graph contrastive learning methods obtain consistent node representations by maximizing the mutual information between multiple views, which can lead to overfitting of the model. We argue that a good graph contrastive learning method should construct structurally heterogeneous but semantically similar multiple views.

To tackle the above problem, we propose a novel Adversarial Alignment and Graph Fusion via Information Bottleneck for Multimodal Emotion Recognition in Conversations, i.e., AGF-IB. Firstly, we use RoBERTa [10], 3D-CNN [11], and OpenSMILE [12] to obtain semantic information in text, video, and audio, respectively. Secondly, we input the extracted three modality features into a multi-layer perceptron (MLP) to map them into separate feature spaces. Thirdly, we build a generator and a discriminator for the three modal features, respectively, and then use adversarial learning to achieve cross-modal feature fusion and eliminate the heterogeneity between different modalities. Fourthly, we construct a new graph contrastive representation learning architecture via information bottleneck (IB), which captures complementary semantic information within and between modalities and intra-class and inter-class differences by performing contrastive representation learning on nodes with different emotions in the same modality and

nodes with the same emotion in different modalities and utilizing IB to minimize the mutual information between multiple views, to obtain a structurally heterogeneous but semantically similar multiple views and more explicit representation of the boundary distribution. Finally, we use MLP for emotion classification.

1.1. Our contributions

Therefore, MERC should not only consider eliminating the heterogeneity among the three modalities of video, audio, and text but also learn how to capture the complementary semantic information within and between modalities and the intra-class and inter-class differences. Inspired by the above analysis, we propose a novel Adversarial Alignment and Graph Fusion via Information Bottleneck for Multimodal Emotion Recognition in Conversations (AGF-IB) to learn better emotion class boundary information. The main contributions of this paper are summarized as follows:

- A novel Adversarial Alignment and Graph Fusion via Information Bottleneck for Multimodal Emotion Recognition in Conversations architecture is present, i.e., AGF-IB. AGF-IB can learn better emotion class boundary information.
- A new cross-modal feature alignment method with adversarial learning is designed to eliminate heterogeneity among modalities.
- A novel graph contrastive representation learning framework via information bottleneck is present to enhance the correlation of intra-modal and inter-modal semantic information, learn the intra-class and inter-class differences, and obtain structurally heterogeneous but semantically similar multiple views.
- Finally, extensive experiments are conducted on two benchmark datasets, i.e., MELD and IEMOCAP. The experimental results show that the emotion recognition effect of AGF-IB is better than the existing comparison algorithms. Furthermore, AGF-IB can be applied to other multimodal tasks in a plug-and-play manner, e.g., humor detection.

The rest of this paper is organized as follows. Section 2 presents the related work of prior MERC. Section 3 describes the multi-modal emotion recognition task and presents the multi-modal data processing flow. Section 4 illustrates the proposed neural network AGF-IB. Section 5 describes the datasets and evaluation metrics used. The related experimental results and discussion on the IEMOCAP and MELD datasets are shown in Section 6. Finally, we conclude our work and illustrate future work.

2. Related work

2.1. Multimodal emotion recognition in conversation

As an interdisciplinary study (e.g., brain science and cognitive science, etc.), MERC has received extensive attention from researchers [13]. The current mainstream MERC research mainly includes sequential context modeling, speaker relationship modeling, and multimodal feature fusion modeling. The sequential context modeling method mainly combines the semantic information of the context to classify the emotion at the current moment. The speaker relationship modeling method mainly extracts the semantic information of the dialogue relationship between speakers through the graph convolution operation. The multimodal feature fusion modeling method mainly achieves cross-modal feature fusion by capturing intra-modal and inter-modal complementary semantic information.

In the modeling method based on sequential context, Poria et al. [14] proposed bidirectional long-short term memory (Bi-LSTM), which can extract contextual semantic information of forward and reverse sequence. However, bc-LSTM has a limited ability to model long-distance context dependencies. In response to the above problems, Beard et al. [15] proposed recursive multi-attention (RM), which uses multi-gated memory units to iteratively update the memory network, thereby realizing the memory of global context information. Although sequential context-based modeling can achieve certain results in emotion recognition, it ignores the intra-modal and inter-modal complementary semantic information.

In the modeling method based on multi-modal feature fusion, Zadeh et al. [6] proposed Tensor Fusion Network (TFN), which maps multi-modal features into three-dimensional space through tensor outer product operation, to realize information interaction between multi-modal features. However, the feature dimension of TFN is high, which is prone to an overfitting effect. To alleviate the problems of TFN, Liu et al. [7] proposed a Low-rank Fusion Network (LFN), which realizes dimensionality reduction of tensors through low-rank decomposition operations and has achieved performance improvement in emotion recognition. Hu et al. [8] proposed Multi-modal Fusion via Deep Graph Convolution Network (MMGCN), which can effectively utilize the complementary semantic information between multi-modal features. Although the above methods can achieve cross-modal feature fusion, they all map the features of different modalities into the same feature space, which makes it challenging to eliminate the heterogeneity between different modalities.

In the modeling method based on speaker relationship, Ren et al. [16] proposed a Latent Relation-Aware Graph Convolutional Network (LR-GCN), which first constructs a speaker relation graph and then introduces a multi-head attention mechanism to capture latent relations between utterances. However, fully connected graphs introduce noise information. Nie et al. proposed [17] Correlation-based Graph Convolutional Network (C-GCN), this method can capture the correlation inter and intra modalities and realize the effective use of multimodal information. Although the modeling method based on speaker relationships can fully use the semantic information of speaker dialogue relationships and cross-modal semantic information, it ignores the differences between different emotion categories.

2.2. Generative adversarial learning

In the field of multimodal emotion recognition in conversations, data imbalance is a common problem, which will lead to biased learning of the model [18]. Therefore, researchers began to use generative adversarial learning to generate new samples that fit the original data distribution. Specifically, previous work generates new samples by minimizing the data distribution learned by the generator and the discriminator.

Su et al. [3] proposed Corpus-Aware Emotional CycleGAN (CAEmo-CyGAN), which innovatively introduces a target-to-source generator to generate new samples that more closely match the original data distribution. CAEmoCyGAN enhances the model's ability to learn unbiased representations. Chang et al. [19] proposed Adversarial Cross Corpora Integration (ACCI), which uses an adversarial autoencoder to generate samples with contextual semantic information and uses emotion labels as auxiliary constraints for the model. Although using new samples generated by generative adversarial learning can effectively alleviate the data imbalance problem, eliminating the heterogeneity between modalities based on GAN is still an open problem.

2.3. Contrastive learning

Self-supervised learning (SL), an essential branch of deep learning (DL), has received increasing research attention because of its powerful ability to learn representations. Contrastive representation learning (CRL) is one of the representative methods for SL. Specifically, CRL learns discriminative features by continuously shrinking the distance (e.g., Euclidean distance and Mahalanobis distance, etc.) between positive samples and expanding the distance between positive and negative samples. Previous work usually obtains representations of features by maximizing the mutual information (MI) between model inputs and learned representations.

Li et al. [20] proposed contrastive predictive coding (CPC) to address the lack of large-scale datasets for emotion recognition tasks. Through unsupervised contrastive representation learning, CPC can learn latent emotional semantic information from unlabeled data. Kim et al. [21] proposed contrastive adversarial learning (CAL) to solve the problem of existing methods relying too much on supervised information. CAL learns complex semantic emotional information by comparing samples with strong emotional features and samples with weaker emotions. Wang et al. [22] designed a new architecture composed of three networks (i.e., FacesNet, SceneNet, and ObjectsNet) to improve the feature fusion ability of the model and solve the problem of missing critical semantic information. Although contrastive representation learning can enhance the representation of emotional information, the above methods ignore the intra-modal and inter-modal information interaction and intra-class and inter-class contrastive representation learning.

3. Preliminary information

In this section, the Multimodal Emotion Recognition in Conversations (MERC) task is defined in mathematical terms. In addition, we also describe the data preprocessing methods of different modalities as follows: (1) Word Embedding: To eliminate the ambiguity of words, this paper uses RoBERTa [10] to obtain the embedding representation of word vectors (2) Visual Feature Extraction: We use 3D-CNN to capture deeper image features in videos and reduce the introduction of noisy information. (3) Audio Feature Extraction: We use OpenSMILE [12] to extract audio signals from different speakers.

3.1. Multimodal feature extraction

The experimental datasets IEMOCAP and MELD in this paper consist of three modalities, which are stored in the form of text, video, and audio, respectively. For the features of different modalities, we use a specific data preprocessing method for feature extraction to obtain feature vector representations with less noise information and rich semantic information. We describe how the features are encoded for each modality as follows.

3.1.1. Word embedding

To disambiguate words and obtain feature vectors with rich semantic information, following previous work [23–25], we use the RoBERTa model [10] to encode words. In this paper, we use sentence-level encoding to encode each utterance of the speaker, and obtain a contextual semantic representation $\varphi_i = \{\varphi_i^1, \varphi_i^2, \dots, \varphi_i^m\}$ containing the entire sentence. Among them, m is the dimension of word embedding. Due to limited computing resources, we only take the first 100-dimensional vectors encoded by the RoBERTa model as our word embedding representation ξ_u .

3.1.2. Visual feature extraction

The speaker's facial expression and behavior reflect his inner emotional state. Therefore, we capture the speaker's facial expressions and action changes from the video frames, thereby extracting semantic information related to the speaker's emotional changes. In this paper, following previous work [11,26,27], we use the 3D-CNN model to obtain a 512-dimensional feature vector ξ_v .

3.1.3. Audio feature extraction

The fluctuation of the voice in the audio signal also reflects the emotional changes in the speaker's heart. Sometimes a person's actions may not truly reflect his emotions, but his tone changes cannot be faked. Therefore, following previous work [11,26,27], we use OpenSMILE to extract the speaker's audio features ξ_a .

3.2. Information bottleneck

Information bottleneck theory (IB) describes two processes during neural network training, i.e., feature fitting and feature compression. IB theory argues that during the training process, the model should maintain task-related information while discarding redundant information that is irrelevant to the task, which can improve the robustness of the model. Formally, for the input data x of the neural network, the label information of the downstream task is y , and the information-compressed feature representation h can be obtained using the IB strategy. The optimization goals of IB are as follows:

$$\max_{\mathbf{H}} I(\mathbf{Y}, \mathbf{H}, \theta) - \beta I(\mathbf{X}, \mathbf{H}, \theta) \quad (1)$$

where β is a scaling factor, θ is a learnable parameter.

For the mutual information $I(Y, H)$ between the label Y and the hidden layer feature H in Eq. (1), we get from the definition of mutual information:

$$\begin{aligned} I(\mathbf{Y}, \mathbf{H}) &= \int dy d\mathbf{h} p(\mathbf{y}, \mathbf{h}) \log \frac{p(\mathbf{y}, \mathbf{h})}{p(\mathbf{y})p(\mathbf{h})} \\ &= \int dy d\mathbf{h} p(\mathbf{y}, \mathbf{h}) \log \frac{p(\mathbf{y} | \mathbf{h})}{p(\mathbf{y})} \end{aligned} \quad (2)$$

where $p(y|h)$ represent the true label distribution under condition h , $q(y|h)$ represent the predicted label distribution under condition h . However, $p(y|h)$ is a difficult estimation problem. Inspired by variational estimation [28], the paper utilizes $q(y|h)$ as the variational approximation of $p(y|h)$, because $p(y|h)$ can be calculated directly.

Since the Kullback–Leibler divergence $KL \geq 0$, we have:

$$\begin{aligned} &\text{KL}[p(\mathbf{y} | \mathbf{h}), q(\mathbf{y} | \mathbf{h})] \geq 0 \\ \Rightarrow &\int dy p(\mathbf{y} | \mathbf{h}) \log \frac{p(\mathbf{y} | \mathbf{h})}{q(\mathbf{y} | \mathbf{h})} \geq 0 \\ \Rightarrow &\int dy p(\mathbf{y} | \mathbf{h}) \log p(\mathbf{y} | \mathbf{h}) \\ &\geq \int dy p(\mathbf{y} | \mathbf{h}) \log q(\mathbf{y} | \mathbf{h}) \end{aligned} \quad (3)$$

Combining Eqs. (2) and (3), we know:

$$\begin{aligned} I(\mathbf{Y}, \mathbf{H}) &\geq \int dy d\mathbf{h} p(\mathbf{y}, \mathbf{h}) \log \frac{q(\mathbf{y} | \mathbf{h})}{p(\mathbf{h})} \\ &= \int dy d\mathbf{h} p(\mathbf{y}, \mathbf{h}) \log q(\mathbf{y} | \mathbf{h}) + H(\mathbf{y}) \\ &\geq \int dy d\mathbf{h} p(\mathbf{y}, \mathbf{h}) \log q(\mathbf{y} | \mathbf{h}) \\ &= \int dy p(\mathbf{y}) \int d\mathbf{h} p(\mathbf{h} | \mathbf{y}) \log q(\mathbf{y} | \mathbf{h}). \end{aligned} \quad (4)$$

For the input data and hidden layer features $I(X, H)$ in Eq. (1), we get:

$$\begin{aligned} I(\mathbf{H}, \mathbf{X}) &= \int d\mathbf{h} d\mathbf{x} p(\mathbf{x}, \mathbf{h}) \log \frac{p(\mathbf{h}, \mathbf{x})}{p(\mathbf{h}), p(\mathbf{x})} \\ &= \int d\mathbf{h} d\mathbf{x} p(\mathbf{x}, \mathbf{h}) \log \frac{p(\mathbf{h} | \mathbf{x})}{p(\mathbf{h})} \end{aligned} \quad (5)$$

Similarly, $p(h)$ is also a difficult estimation problem. The paper utilizes $r(z)$ as the variational approximation of $p(h)$. Since $\text{KL}[p(\mathbf{h}), r(\mathbf{h})] \geq 0 \Rightarrow \int d\mathbf{h} p(\mathbf{h}) \log p(\mathbf{h}) \geq \int d\mathbf{h} p(\mathbf{h}) \log r(\mathbf{h})$, we get an upper bound:

$$\begin{aligned} I(\mathbf{H}, \mathbf{X}) &\leq \int d\mathbf{x} d\mathbf{z} p(\mathbf{x}) p(\mathbf{h} | \mathbf{x}) \log \frac{p(\mathbf{h} | \mathbf{x})}{r(\mathbf{h})} \\ &= \int d\mathbf{x} p(\mathbf{x}) \int d\mathbf{h} p(\mathbf{h} | \mathbf{x}) \log \frac{p(\mathbf{h} | \mathbf{x})}{r(\mathbf{h})}. \end{aligned} \quad (6)$$

According to the derivation process of the above equation, we obtain the lower bound of the information bottleneck theory as follows:

$$\begin{aligned} I(\mathbf{Y}, \mathbf{H}) - \sum_{v=1}^V \beta I(\mathbf{H}, \mathbf{X}) \\ \geq \int dy p(\mathbf{y}) \int d\mathbf{h} p(\mathbf{h} | \mathbf{y}) \log q(\mathbf{y} | \mathbf{h}) \\ - \beta \int d\mathbf{x} p(\mathbf{x}) \int d\mathbf{h} p(\mathbf{h} | \mathbf{x}) \log \frac{p(\mathbf{h} | \mathbf{x})}{r(\mathbf{h})} \end{aligned} \quad (7)$$

Research has proven that the upper bound of the information bottleneck theory is equivalent to the InfoNCE loss [28].

4. Methodology

4.1. Task definition

The task of Multimodal Emotion Recognition in Conversation (MERC) aims to predict the emotion label of each utterance from a conversation containing textual, acoustic and visual modalities. The goal of MERC is to determine the emotional state expressed by each utterance from predefined emotion categories by comprehensively considering the textual content, sound characteristics and visual information of the utterance. Specifically, in MERC, a conversation is viewed as consisting of N consecutive utterances $\{u_1, u_2, \dots, u_N\}$ and M speakers $\{s_1, s_2, \dots, s_M\}$, and each utterance is uttered by a specific speaker in the conversation. In addition, utterance u_i includes information of different modalities, such as text content, voice characteristics, and speaker's facial expressions. We represent textual, acoustic, and visual modality sequences of all utterances in the conversation as $[\xi_u^1; \xi_u^2; \dots; \xi_u^N] \in \mathbb{R}^{N \times d_u}$, $[\xi_a^1; \xi_a^2; \dots; \xi_a^N] \in \mathbb{R}^{N \times d_a}$, and $[\xi_v^1; \xi_v^2; \dots; \xi_v^N] \in \mathbb{R}^{N \times d_v}$, respectively, where d_u is the text dimension, d_a is the audio dimension, and d_v is the video dimension. By combining different modal information, the model can more comprehensively understand and identify the emotional changes and expressions contained in the conversation.

4.2. The design of the AGF-IB structure

To increase the performance of multi-modal emotion recognition, we propose a novel Adversarial Alignment and Graph Fusion via Information Bottleneck for Multimodal Emotion Recognition in Conversations, namely AGF-IB. The overall architecture of AGF-IB is shown in Fig. 2.

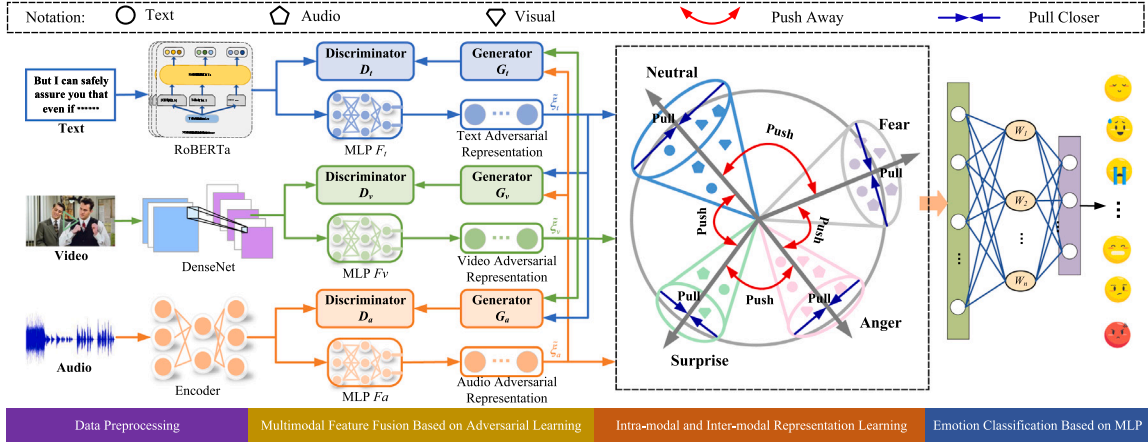


Fig. 2. The overall framework of the Adversarial Alignment and Graph Fusion via Information Bottleneck consists of a data preprocessing layer, a multimodal feature fusion layer, a graph contrastive representation learning layer, and an emotion classification layer.

4.2.1. TGAN: Tri-modal Generative Adversarial Networks

Multimodal features provide more emotional semantic information for the MERC task. However, multimodal data are heterogeneous and noisy, which makes cross-modal feature fusion difficult. Therefore, to effectively eliminate the heterogeneity between modalities and make full use of the complementary semantic information of multimodalities, we design a Tri-modal Generative Adversarial Networks (TGAN) to eliminate the data distribution differences between different modal features.

Specifically, firstly, we use MLP to dimensionally align the three modality features and map them into three separate feature spaces. The formulas are as follows:

$$\begin{aligned}\tilde{\xi}_u &= F_t(\xi_u) \in \mathbb{R}^d \\ \tilde{\xi}_v &= F_v(\xi_v) \in \mathbb{R}^d \\ \tilde{\xi}_a &= F_a(\xi_a) \in \mathbb{R}^d\end{aligned}\quad (8)$$

where d denotes the dimension that maps the three modal features to a separate representation space. $F_t(\cdot)$, $F_v(\cdot)$, $F_a(\cdot)$ represents the MLP layer.

Secondly, we build a text generator and a text discriminator. The input to the text generator is audio features $\tilde{\xi}_a$ and video features $\tilde{\xi}_v$. The input of the text discriminator is the fused features generated by the text generator containing three modal information. The objective optimization functions for the text generator and discriminator are as follows:

$$\begin{aligned}\min_{G_t} \mathcal{L}_{Gen}(G_t, D_t) &= \mathbb{E}_{\tilde{\xi}_a \sim P_{data}}(\tilde{\xi}_a) [\log(1 - D_t(G_t(\tilde{\xi}_a)))] \\ &+ \mathbb{E}_{\tilde{\xi}_v \sim P_{data}}(\tilde{\xi}_v) [\log(1 - D_t(G_t(\tilde{\xi}_v)))] \\ \max_{D_t} \mathcal{L}_{Dis}(G_t, D_t) &= \mathbb{E}_{T \sim P_{data}}(T) [\log D_t(T)] \\ &+ \mathbb{E}_{\tilde{\xi}_a \sim P_{data}}(\tilde{\xi}_a) [\log(1 - D_t(G_t(\tilde{\xi}_a)))] \\ &+ \mathbb{E}_{\tilde{\xi}_v \sim P_{data}}(\tilde{\xi}_v) [\log(1 - D_t(G_t(\tilde{\xi}_v)))]\end{aligned}\quad (9)$$

where G_t and D_t represent text generator and text discriminator, $\tilde{\xi}_a \sim P_{data}$ represents sampling samples from the data that conforms to the audio feature distribution law, and $\tilde{\xi}_v \sim P_{data}$ represents sampling samples from the data that conforms to the video feature distribution law.

Thirdly, we build an audio generator and an audio discriminator. The input to the audio generator is text features and video features. The input of the audio discriminator is the fused features generated by the audio generator containing three modal information. The objective

optimization functions for the audio generator and discriminator are as follows:

$$\begin{aligned}\min_{G_a} \mathcal{L}_{Gen}(G_a, D_a) \\ \max_{D_a} \mathcal{L}_{Dis}(G_a, D_a)\end{aligned}\quad (10)$$

where G_a and D_a represent audio generator and audio discriminator.

Finally, we build a video generator and a video discriminator. The input of the video generator is text features and audio features. The input of the video discriminator is the fused features generated by the video generator containing three modal information. The objective optimization functions for the video generator and discriminator are as follows:

$$\begin{aligned}\min_{G_v} \mathcal{L}_{Gen}(G_v, D_v) \\ \max_{D_v} \mathcal{L}_{Dis}(G_v, D_v)\end{aligned}\quad (11)$$

where G_v and D_v represent video generator and video discriminator.

It should be noted that after training the three-modal generative confrontation network, we proceed to the training of subsequent tasks.

4.2.2. Speaker relation graph construction

We use a graph structure to extract semantic information of speaker dialogue relations. Specifically, we construct a directed graph of speaker relations $\mathcal{G}_M = \{\mathcal{V}_M, \mathcal{E}_M, \mathcal{R}_M, \mathcal{W}_M\}$ for the three modal features of video, audio and text respectively, where $M \in \{T, V, A\}$, the node v_i^M ($v_i^M \in \mathcal{V}_M$) is composed of unimodal features (i.e., $\tilde{\xi}_a, \tilde{\xi}_v, \tilde{\xi}_u$), the directed edge r_{ij}^M ($r_{ij}^M \in \mathcal{E}_M$) indicates that there is a dialogue relationship between the node v_i^M and the node v_j^M , and ω_{ij}^M ($\omega_{ij}^M \in \mathcal{W}_M, 0 \leq \omega_{ij}^M \leq 1$) is the weight of the edge r_{ij}^M , and $r^M \in \mathcal{R}_M$ is the edge type. In particular, in the MERC task, we follow previous work [8,16,27] to construct a fully connected dialogue graph, i.e., nodes are all connected within the context window. Furthermore, there is only one type of edge in the graph, i.e., dialogue relationship. Since the computational complexity of GCN is $O(n^2)$, this leads to high computational resources required. Therefore, we set the context window size to 10.

To capture the key semantic information in the nodes, we use the attention mechanism to calculate the weight of the edge, and perform information aggregation according to the edge weight. Firstly, we use MLP to dynamically learn the correlation between node i and node j . The formula is defined as follows:

$$\epsilon_{ij}^M = W_{\theta_1}^M \left(\text{GELU} \left(W_{\theta_2}^M \left[\tilde{\xi}_i^M \oplus \tilde{\xi}_j^M \right] \right) \right) \quad (12)$$

where $W_{\theta_1}^M$, $W_{\theta_2}^M$ are learnable network parameters, and \oplus represents the vector concatenation operation.

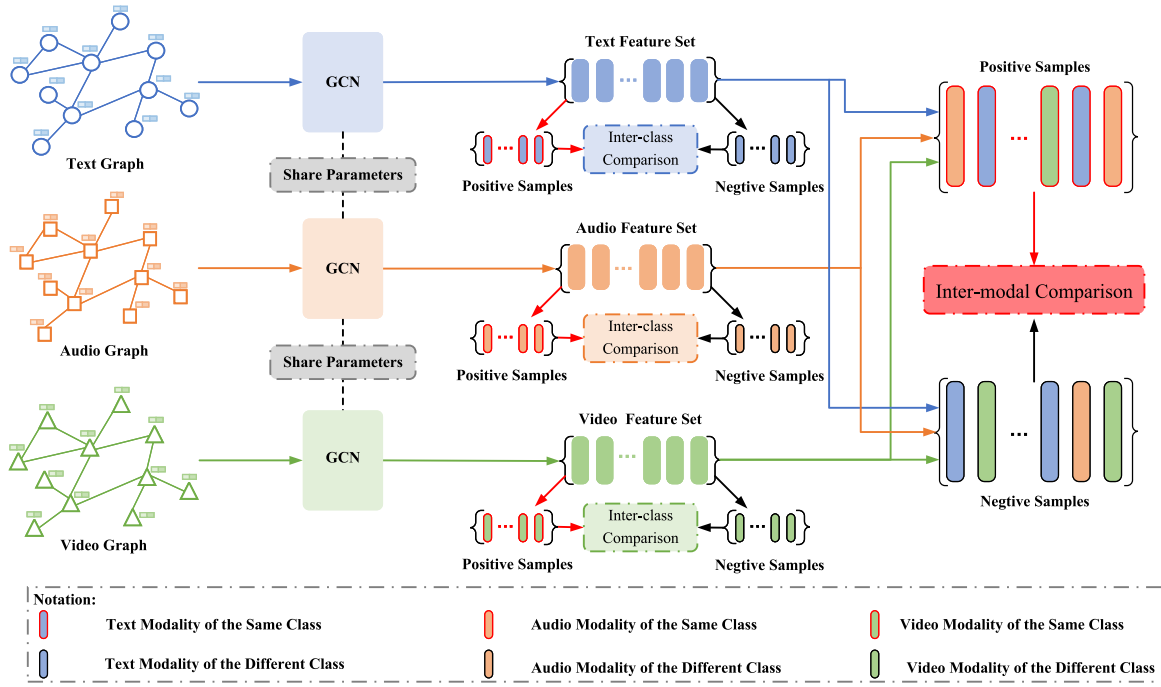


Fig. 3. The overall process of graph contrastive representation learning includes intra-modal and inter-modal comparison, and intra-class and inter-class comparison.

Secondly, we use a softmax function to normalize the correlation between node i and node j to obtain the attention score for each edge. The formula is defined as follows:

$$\omega_{ij}^M = \text{softmax}(\varepsilon_{ij}^M) = \frac{\exp(\varepsilon_{ij}^M)}{\sum_{\eta \in \mathcal{N}_i} \exp(\varepsilon_{i\eta}^M)} \quad (13)$$

where \mathcal{N}_i represents the first-order neighbor nodes of node i . The larger ω_{ij}^M represents the stronger correlation between node i and node j .

Finally, we update the node representations using a GCN followed by a GELU activation function. The formula for GCN encoding is as follows:

$$\psi_i^M(t) = \text{GELU} \left(\sum_{r \in R} \sum_{j \in \mathcal{N}_i^r} \frac{1}{|\mathcal{N}_i^r|} \left(\omega_{ij}^M W_{\theta_1}^M \psi_j^M(t-1) + \omega_{ii}^M W_{\theta_2}^M \psi_i^M(t-1) \right) \right) \quad (14)$$

where \mathcal{N}_i^r is the set of first-order neighbor nodes of node i under the edge relationship $r \in R$, $|\mathcal{N}_i^r|$ is the modulus of \mathcal{N}_i^r , and $\psi_i^M(t)$ is the feature vector encoded by GCN.

4.3. IB loss and mutual information estimation

For the given input data set $X = \{x_1, x_2, \dots, x_N\}$, it contains one positive sample from $p(x_{t+k} | c_t)$ and $N-1$ positive samples from $p(x_{t+k})$ negative sample, the InfoNCE loss is defined as follows:

$$\mathcal{L}_N = -\mathbb{E}_X \left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)} \right] \quad (15)$$

where c_t represents the contextual information, i.e., $c_t = \{x_1, x_2, \dots, x_t\}$.

However, the sample x_t should be derived from the conditional distribution $p(x_t + k | c_t)$, not $p(x_t + k)$. The conditional distribution

is derived as follows:

$$\begin{aligned} p(d = i | X, c_t) &= \frac{p(x_i | c_t) \prod_{l \neq i} p(x_l)}{\sum_{j=1}^N p(x_j | c_t) \prod_{l \neq j} p(x_l)} \\ &= \frac{p(x_i | c_t)}{p(x_i)} \\ &= \frac{\sum_{j=1}^N \frac{p(x_j | c_t)}{p(x_j)}}{\sum_{j=1}^N \frac{p(x_j | c_t)}{p(x_j)}} \end{aligned} \quad (16)$$

Minimizing the InfoNCE loss \mathcal{L}_N is equivalent to maximizing the lower bound of mutual information. Inspired by the InfoNCE loss, minimizing mutual information is equivalent to optimizing negative InfoNCE. Formally, nodes of the same type are regarded as positive sample pairs (i.e., $\{(e_i, \tilde{e}_i) | v_i \in \mathcal{V}_M\}$), while nodes of different types are regarded as negative sample pairs (i.e., $\{(e_i, \tilde{e}_j) | v_i, v_j \in \mathcal{V}_M, i \neq j\}$).

$$I(\mathbf{E}, \tilde{\mathbf{E}}) = \sum_{v_i \in \mathcal{V}} \log \frac{\exp(s(\mathbf{e}_i, \tilde{\mathbf{e}}_i) / \tau)}{\sum_{v_j \in \mathcal{V}} \exp(s(\mathbf{e}_i, \tilde{\mathbf{e}}_j) / \tau)} \quad (17)$$

Where \mathbf{E} represents the original view and $\tilde{\mathbf{E}}$ represents the augmented view. $s(\cdot)$ is used to calculate the similarity between nodes.

4.4. IMCL: Intra-modal and inter-modal contrastive learning via IB

IMCL aims to learn complementary semantic information between modalities and obtain a more discriminative embedding representation through contrastive learning method. Different from existing graph contrastive learning methods, we use information bottleneck theory to minimize the mutual information between multiple views. Specifically, in IMCL, positive samples are represented by samples of the same class in the same modality, while negative samples are represented by samples of the same class in different modalities. The intra-modal and inter-modal contrastive loss is defined as follows:

$$\mathcal{L}_{IMCL} = -\mathbb{E}_s \left[\frac{\sum_{i=1}^N \exp(P_i^M / \tau)}{\sum_{i=1}^{2N} \sum_{j=1}^{2M} \exp(P_i^M + Q_j^M / \tau)} \right] \quad (18)$$

where $P_i^M = s(\mu^M, \chi_i)$, $Q_j^M = s(\mu^M, \delta_j^M)$, μ^M denotes the anchor embedded representation, N denotes the number of positive samples, M denotes the number of negative samples, χ_i^M and δ_j^M denote the embedded representations of positive and negative samples, respectively. It should be noted that χ_i^M and δ_j^M are the same modality and different classes.

However, if Eq. (18) is used as a contrastive loss, the model may fall into a local optimal solution. i.e., $s(\mu^M, \chi_i^M)$ can be minimized but $s(\mu^M, \delta_j^M)$ cannot be maximized. The above situation is because when the similarity between negative sample pairs is 0, no matter how much the similarity between positive sample pairs is, the contrastive loss of the model tends to the minimum value. Our desired goal is that $s(\mu^M, \chi_i^M)$ can be minimized and $s(\mu^M, \delta_j^M)$ can be maximized. Therefore, we introduce a regularization term to ensure that the similarity between positive sample pairs can be maximized and the similarity between negative sample pairs can be minimized. The formula is defined as follows:

$$\mathcal{L}_{IMCL}^R = \mathbb{E}_S \left[\frac{1}{2N} \sum_{i=1}^N \left\| s(\mu^M, \delta_j) - \beta \right\|^2 \right] \quad (19)$$

$$\tilde{\mathcal{L}}_{IMCL} = \mathcal{L}_{IMCL} + \mathcal{L}_{IMCL}^R$$

where \mathcal{L}_{IMCL} is the regularization loss for IMCL, β is a hyperparameter. IMCL encourages high similarity between samples of the same class in the same modality, and forces low similarity between samples of the same class in different modalities. The overall process of IMCL and ICCL is shown in Fig. 3.

4.5. ICCL: Intra-class and Inter-class Contrastive Learning via IB

Similar to IMCL, ICCL aims to learn intra-class and inter-class semantic information with differences through contrastive learning. Specifically, the intra-class and inter-class contrastive loss is defined as follows:

$$\mathcal{L}_{ICCL} = -\mathbb{E}_S \left[\frac{\sum_{i=1}^N \exp(P_i^M/\tau)}{\sum_{i=1}^N \sum_{j=1}^M \exp(T_i^M + Q_j^M/\tau)} \right] \quad (20)$$

where $T_i^M = s(\mu_i^M, \chi_i^M)$, χ_i^M and δ_j^M belong to samples of the same modality. Similar to IMCL, we also introduce regularization terms to strengthen the similarity between positive sample pairs and reduce the similarity between negative samples. The formula is defined as follows:

$$\mathcal{L}_{ICCL}^R = \mathbb{E}_S \left[\frac{1}{N} \sum_{i=1}^N \left\| s(\mu^M, \chi_i^M) - 1 \right\|^2 \right] \quad (21)$$

$$\tilde{\mathcal{L}}_{ICCL} = \mathcal{L}_{ICCL} + \mathcal{L}_{ICCL}^R$$

where \mathcal{L}_{ICCL}^R is the regularization loss for ICCL.

To understand why IMCL and ICCL are effective, we introduce two important theories in contrastive learning, i.e., alignment and uniformity. Specifically, alignment is used to measure the spatial distance between positive pairs, and the formula is defined as follows:

$$\ell_{\text{ali}}(f; \alpha) \triangleq \mathbb{E}_{(x,y) \sim p_{\text{pos}}} \left[\|f(x) - f(y)\|_2^2 \right], \quad \alpha > 0 \quad (22)$$

where p_{pos} represents the spatial distribution between positive pairs. The goals of Eq. (22) are very consistent with those of contrastive learning. Similarly, for IMCL and ICCL, the alignment metric is defined as follows:

$$\ell_{\text{ali}}(f; \alpha) \triangleq \mathbb{E}_{(x,y) \sim p_{\text{pos}}} \left[\|f(x, \theta) - f(y, \theta')\|_2^2 \right] \quad (23)$$

The consistency is defined as follows:

$$\ell_{\text{uni}}(f; \alpha) \triangleq \log \mathbb{E}_{(x,y) \sim p_{\text{pos}}} \left[e^{-\alpha \|f(x;\theta) - f(y;\theta)\|_2^2} \right] \quad (24)$$

After analyzing IMCL and ICCL, they can achieve better alignment while improving uniformity.

Table 1

The division of training set, validation set and test set of IEMOCAP and MELD data sets.

Datasets	Partition	Utterance Count	Dialogue Count
IEMOCAP	train+val	5810	120
	test	1623	31
MELD	train+val	11 098	1153
	test	2610	280

4.6. Emotion inference subnetwork

After the multimodal feature vectors pass through the matching attention layer, each contextual utterance can be represented as a multimodal fusion vector z^f . We use a multi-layer perceptron deep neural network called the Emotion Inference Subnetwork \mathcal{G}_s with weights W conditioned on z^f . The multi-layer perceptron (MLP) consists of two fully connected layers with ReLU activation functions and connects them to a decision layer. The maximum likelihood function of the Emotion Inference Subnetwork \mathcal{G}_s is defined as follows, where φ is the label of emotion prediction:

$$\mathcal{L}_{CLS} = \arg \max_{\varphi} p(\varphi | z^f; W) = \arg \max_{\varphi} \mathcal{G}_s(z^f, W) \quad (25)$$

where \mathcal{L}_{CLS} is emotion classification loss for the model. The smaller \mathcal{L}_{CLS} , the better the emotion classification effect.

4.7. Model training

The intra-modal and inter-modal, and intra-class and inter-class contrastive losses are obtained by weighted summation of IMCL and ICCL. The formula is defined as follows:

$$\mathcal{L}_{\text{hybrid}} = \lambda \tilde{\mathcal{L}}_{IMCL} + (1 - \lambda) \tilde{\mathcal{L}}_{ICCL} \quad (26)$$

The overall loss for model training is obtained by summing the classification loss and the contrastive loss. The formula for the model training loss is defined as follows:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{CLS} + \mathcal{L}_{\text{hybrid}} \quad (27)$$

where $\mathcal{L}_{\text{overall}}$ is the overall loss of the model. The smaller $\mathcal{L}_{\text{overall}}$, the better the training effect of the model.

5. Experiments

5.1. Benchmark dataset used

The MELD [29] and IEMOCAP [30] multimodal conversation datasets are often used for comparative experiments in MERC. We introduce the situation of the two datasets as follows and show the division of the IEMOCAP and MELD datasets in Table 1.

The Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) contains three modalities, namely video, audio, and text. Therefore, IEMOCAP is a multimodal dataset, and the use of multimodal emotion recognition in conversations methods can enhance the prediction effect of the model. A total of 10 actors and actresses are included in the IEMOCAP dataset, and they communicate in an interactive way. For each conversation, it is annotated by multiple emotion experts, avoiding the subjectivity of human annotation. In addition, the IEMOCAP dataset contains a total of six emotions, namely ‘‘sad’’, ‘‘happy’’, ‘‘angry’’, ‘‘neutral’’, ‘‘frustrated’’ and ‘‘excited’’.

The Multi-modal EmotionLines Dataset (MELD) is also a multi-modal dataset whose corpus consists of dialogues from the TV series Friends. Similar to the IEMOCAP dataset, each conversation is also annotated by multiple emotion experts. In addition, the MELD dataset contains a total of seven emotions, namely ‘‘disgust’’, ‘‘anger’’, ‘‘joy’’, ‘‘fear’’, ‘‘sadness’’, ‘‘neutral’’, and ‘‘surprise’’.

Table 2

Experimental results with our method and other baseline on IEMOCAP dataset. The best result in each column is in bold. Average(w) represents the weighted average.

Methods	IEMOCAP													
	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average(w)	
	Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1	WAA	WF1	
TextCNN	27.73	29.81	57.14	53.83	34.36	40.13	61.12	52.47	46.11	50.09	62.94	55.78	48.93	48.17
bc-LSTM	29.16	34.49	57.14	60.81	54.19	51.80	57.03	56.75	51.17	57.98	67.12	58.97	55.23	54.98
bc-LSTM+Att	30.56	35.63	56.73	62.09	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41	56.32	56.19
DialogueRNN	25.63	33.11	75.14	78.85	58.56	59.24	64.76	65.23	80.27	71.85	61.16	58.97	63.42	62.74
DialogueGCN	40.63	42.71	89.14	84.45	61.97	63.54	67.51	64.14	65.46	63.08	64.13	66.90	65.21	64.14
CT-Net	47.97	51.36	78.01	79.94	69.08	65.82	72.98	67.21	85.35	78.74	52.27	58.83	68.01	67.55
LR-GCN	54.24	55.51	81.67	79.14	59.13	63.84	69.47	69.02	76.37	74.05	68.26	68.91	68.52	68.35
MM-DFN	40.17	42.22	74.27	78.98	69.13	66.42	70.25	69.97	76.99	75.56	68.58	66.33	68.21	68.18
MMIM	33.17	38.97	79.90	71.62	63.03	58.26	61.88	67.20	77.54	75.68	63.68	65.39	63.89	64.39
GCNet	40.85	48.74	74.67	72.15	63.81	61.93	60.91	65.40	84.52	77.31	64.59	63.65	65.37	65.90
ARGF	26.39	–	68.98	–	54.95	–	62.35	–	64.21	–	68.50	–	60.20	59.81
M2FNet	65.92	60.00	79.18	82.11	65.80	65.88	75.37	68.21	74.84	72.60	66.87	68.31	69.69	69.86
AGF-IB	71.88	69.96	74.64	81.58	67.25	63.80	73.79	68.37	82.66	79.15	60.45	63.95	70.46	70.36

5.2. Implementation details

In this section, we describe the implementation details of the model during training. We divide the benchmark dataset into three parts. The first part is the training set for model training, the second part is the validation set for updating the network parameters, and the third part is the test set for evaluating the emotional prediction effect of the model. The experimental environment of this paper is the Windows 10 operating system, and the hardware driver is a computer with Nvidia RTX 3090. We use Python 3.8, and Pytorch 1.9.1 version to complete the construction of deep learning algorithms. To ensure the effective convergence of the model, this paper uses the highly stable Adam algorithm [31] to optimize the network parameters. In addition, during the experiment, we set the epochs size to 60, batch size to 32, learning rate to 0.0005, dropout to 0.5, and weight decay coefficient to 0.00001. If not specified otherwise, we use RoBERTa-Large by default to extract contextual semantic information of text features.

5.3. Evaluation metrics

To compare the emotion recognition effect of our algorithm and other baseline algorithms, we use four evaluation metrics: (1) Accuracy; (2) F1; (3) Weighted average accuracy (WAA); (4) Weighted average F1 (WAF1).

5.4. Baseline models

We do extensive comparative experiments on two popular datasets to count the emotion recognition effect of the model proposed in this paper. Some recent comparison algorithms are described below:

TextCNN: The TextCNN proposed by Kim et al. [32] uses Convolutional Neural Networks (CNN) for emotion recognition of dialogues. TextCNN exploits the local attention mechanism of convolution kernels to extract contextual utterances with emotional polarity in texts. However, TextCNN cannot model the context of long-range dependencies and can only model unimodal features.

bc-LSTM: The bidirectional LSTM (bc-LSTM) proposed by Poria et al. [14] can not only model long-range contextual dependencies, but also extract contextual information in two opposite directions, thereby eliminating word ambiguity. However, bc-LSTM does not model speaker relations.

DialogueRNN: The DialogueRNN proposed by Ghosal et al. [11] consists of three gated recurrent units (i.e., global GRU, party GRU and emotion GRU), which are able to distinguish between speakers.

DialogueGCN: DialogueGCN proposed by Ghosal et al. [27] is the first to use graph convolutional neural networks (GCNs) to model speaker relations. DialogueGCN simulates the dialogue relationship

between speakers by constructing a fully connected directed graph, which can fuse the contextual semantic information and the semantic information of the dialogue relationship between speakers. However, the fully connected graph constructed by DialogueGCN may introduce noisy information.

CTNet: The Conversational Transformer Network proposed by Lian et al. [33] comprehensively considers intra-modal and inter-modal modeling, and captures long-range contextual information by using a cross-modal Conversational Transformer architecture.

LR-GCN: The LR-GCN proposed by Ren et al. [16] not only utilizes GCN to model the relationship between speakers, but also utilizes a multi-head attention mechanism to model the latent relationship between utterances. In addition, to speed up the convergence of the model, LR-GCN also introduces a residual structure to transfer more gradient information. LR-GCN has achieved good experimental results.

MM-DFN: Chudasama et al. [24] combines a multi-head attention mechanism with an adaptive triplet loss to learn emotion-related features.

M2FNet: Hu et al. [34] used Multi-modal Fusion Network to dynamically fuse context information and make full use of complementary semantic information between multi-modal features.

MMIM: Han et al. [35] uses Multimodal Infomax to maximize the mutual information (MI) between the maximum single-modal information and multi-modal fusion information to retain the task-related semantic information.

ARGF: Mai et al. [36] a novel adversarial encoder framework to learn unchanged feature information between different modular characteristics. ARGF's goal is to learn the original distribution of data through the encoder.

GCNet: Lian et al. [37] constructs Speaker GNN and Temporal GNN to capture the time and speaker dependence of the context and speaker at the same time.

6. Results and discussion

6.1. Comparison with baselines

This paper compares our proposed emotion recognition algorithm AGF-IB with other deep learning algorithms. Tables 2 and 3 show the recognition accuracy and F1 value of all algorithms on each emotion category on the IEMOCAP and MELD datasets, and the average accuracy and F1 value of the model. Experimental results demonstrate the superiority of our algorithm.

IEMOCAP: As shown in Table 2, AGF-IB has the best emotion recognition effect on the IEMOCAP dataset, and the WAA and WF1 values are 70.46% and 70.36%, respectively. In addition, AGF-IB has the highest accuracy rate on the “happy” classes, and the highest F1 value on the “happy” and “excited” classes, while the accuracy and F1 value of other

Table 3

Experimental results with our method and other baseline on MELD dataset. The best result in each column is in bold. Average(w) represents the weighted average.

Methods	MELD															
	Neutral		Surprise		Fear		Sadness		Joy		Disgust		Anger		Average(w)	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	WAA	WF1
TextCNN	76.23	74.91	43.35	45.51	4.63	3.71	18.25	21.17	46.14	49.47	8.91	8.36	35.33	34.51	56.35	55.01
bc-LSTM	78.45	73.84	46.82	47.71	3.84	5.46	22.47	25.19	51.61	51.34	4.31	5.23	36.71	38.44	57.51	55.94
bc-LSTM+Att	70.45	75.55	46.43	46.35	0.00	0.00	21.77	16.27	49.30	50.72	0.00	0.00	41.77	40.71	58.51	55.84
DialogueRNN	72.12	73.54	54.42	49.47	1.61	1.23	23.97	23.83	52.01	50.74	1.52	1.73	41.01	41.54	56.12	55.97
CT-Net	75.61	77.45	51.32	52.76	5.14	10.09	30.91	32.56	54.31	56.08	11.62	11.27	42.51	44.65	61.93	60.57
MMIM	61.43	75.33	71.52	52.91	0.00	0.00	66.65	7.27	70.18	50.79	0.00	0.00	60.75	46.51	62.80	56.50
GCNet	60.50	74.32	75.00	46.94	0.00	0.00	66.67	53.74	65.37	59.43	0.00	0.00	66.67	38.68	62.18	60.39
M2FNet	72.88	67.98	52.76	58.66	5.57	3.45	50.09	47.03	58.49	55.50	17.69	15.24	57.33	55.25	63.85	62.71
AGF-IB	81.10	81.19	56.16	57.24	6.90	5.06	47.41	37.32	65.92	65.92	2.94	2.94	42.28	45.22	64.14	64.01

categories are slightly lower than other comparison algorithms. The reason is that AGF-IB comprehensively considers the heterogeneity of modalities, the intra-modal and inter-modal complementary semantic information, and the intra-class and inter-class differences. The emotion recognition effect of M2FNet is second, and the values of WAA and WF1 are 69.69% and 69.86%, respectively. The reason why M2FNet is less effective than AGF-IB is that it ignores the heterogeneity of modalities, which leads to poor learning effect of subsequent class boundaries. The effects of other algorithms are relatively poor, and they do not consider the heterogeneity of modalities and the intra-class and inter-class differences. Specifically, the reasons for the performance of AGF-IB better than MMIM may attribute to reserving semantic information related to task-related semantic information by introducing information bottlenecks theory. Compared with some GCN-based methods (e.g., DialogueGCN and GCNet), AGF-IB introduces a contrastive learning mechanism in GCN to construct a multi-view with structural heterogeneity and similar semantics to improve the generalization ability of the model. Compared with the adversarial learning method ARGF, AGF-IB builds a generator and a discriminator for each modal feature separately, and inputs the generated audio features into the text and video discriminators respectively to eliminate the gap between modalities while retaining modality-specific information.

MELD: As shown in Table 3, AGF-IB has the best emotion recognition effect on the MELD dataset, and the WAA and WF1 values are 64.14% and 64.01%, respectively. In addition, AGF-IB has the highest accuracy on the “neutral”, “surprise”, “fear”, “joy”, and “sadness” categories, the F1 values on the “neutral”, “surprise”, “joy”, “sadness”, and “angry” categories are the highest, while the accuracy and F1 values in other categories are slightly lower than other comparison algorithms. In the “fear” and “disgust” categories, the recognition accuracy and F1 value of AGF-IB and other models are low, because the MELD dataset has a serious category imbalance problem.

The analysis of the above experimental results illustrates the superior performance of AGF-IB, which can effectively learn the class boundary information of emotions.

6.2. Importance of the modalities

Since different modal features contain different semantic information, we explored the emotion recognition effect of different modal features on the IEMOCAP and MELD datasets. As shown in Table 4, text features perform best in emotion recognition in single-modal experiments, with WA values of 65.4% and 60.8%, and WF1 values of 60.8% and 60.1% in IEMOCAP and MELD datasets, respectively. We think this is because text is the most direct way for speakers to express their emotions, and it contains the least noisy information. Audio features perform second best for emotion recognition, while video features perform the worst. We think this is because video features contain too much noise information, and it is difficult for the model to extract key information. The emotion recognition effect of the combination of text, audio and video features is the best in all experiments, because

Table 4

The effect of AGF-IB on IEMOCAP and MELD datasets using unimodal features and multimodal features, respectively. T, V, and A represent text, video, and audio modality features. The best result in each column is in bold.

Modality	IEMOCAP		MELD	
	WA	WF1	WA	WF1
T	65.4	64.9	60.8	60.1
A	62.3	62.0	58.6	57.7
V	56.2	54.5	56.8	54.3
T+A+V	70.5	70.4	64.1	64.0

Table 5

Emotion recognition effects of different multimodal feature fusion methods on IEMOCAP and MELD datasets. We use multi-modal features for each method. The best result in each column is in bold.

Methods	IEMOCAP		MELD	
	WA	WF1	WA	WF1
Add	55.2	55.0	57.5	55.9
Concatenate	58.3	57.4	58.6	57.1
Tensor Fusion	63.2	63.0	59.6	58.7
Low-rank Fusion	63.8	63.6	60.8	59.5
Cross-modal Fusion(Ours)	70.5	70.4	64.1	64.0

the model effectively utilizes the complementary semantic information between modalities. The above experimental phenomena also prove the rationality of our mode fusion layer design.

6.3. Effectiveness of cross-modal feature fusion

In this section, to compare the difference between our proposed multimodal feature fusion method and other methods in multi-modal emotion recognition, we compare our method combining trimodal generative adversarial networks and graph contrastive learning with the other four feature fusion methods.

Add: The Add method combines the feature vectors by summing the multimodal features, which ignores the information interaction between the multimodal features.

Concatenate: The Concatenate method is a splicing operation of multi-modal features, which does not model multi-modal features within and between modalities.

TFN: TFN method models the fusion between multimodal features through tensor outer product operations.

LFM: LFM fuses multimodal features through low-rank tensors.

As shown in Table 5, compared with other fusion methods, our cross-modal fusion method achieves the best emotion recognition performance, with WA values of 70.5% and 64.1% and WF1 values of 70.4 and 64.0% on IEMOCAP and MELD datasets, respectively. Specifically, our method improves the WA value by 15.3% and 6.6% and the WF1 value by 15.4% and 8.1% over the Add method on the IEMOCAP and MELD datasets, respectively. This is because the Add method cannot

Table 6

The results of the equal parameters experiment on IEMOCAP and MELD datasets. The parameters of methods with \diamond are incremented to be the same as methods with AGF-IB. The best result in each column is in bold.

Method	Params	IEMOCAP		MELD	
		WAA	WF1	WAA	WF1
bc-LSTM	0.53M	55.2	54.9	57.1	56.4
bc-LSTM \diamond	14.68M	52.9	52.7	53.3	52.9
DialogueRNN	13.19M	63.4	62.7	56.1	56.0
DialogueRNN \diamond	14.68M	62.7	62.2	54.8	54.1
DialogueGCN	12.78M	65.2	64.1	54.9	54.7
DialogueGCN \diamond	14.68M	63.8	62.7	54.4	53.0
AGF-IB	14.68M	70.5	70.4	64.1	64.0

eliminate the heterogeneity among modalities and cannot utilize complementary semantic information between modalities. Compared with the Concatenate method, our method improves the WA value by 12.2% and 6.9%, and the WF1 value by 5.5% and 6.9%, respectively. Similar to the Add method, the Concatenate method cannot take advantage of complementary semantic information between modals. Compared with the Add and Concatenate methods, the Tensor Fusion and Low-rank Fusion methods have significantly improved results, because they utilize complementary semantic information between modalities, and the Low-rank Fusion method can reduce redundant information between modalities. However, they cannot eliminate the heterogeneity between modalities. The above experiments illustrate the superiority of our designed cross-modal approach and the necessity of eliminating modality heterogeneity.

6.4. Equal parameter experiments

To illustrate that our method AGF-IB does not improve the performance of the model due to the increase in the number of parameters, but the performance improvement caused by the architecture design of the model, we conducted experiments with equal parameters on the IEMOCAP and MELD datasets. As shown in Table 6, the WA and WF1 values of emotion recognition of the bc-LSTM, DialogueRNN and DialogueGCN models decreased when the number of parameters increased. In addition, in the process of observing the model training, we find that the model is more prone to overfitting as the number of parameters increases. Therefore, the above phenomenon shows that our model architecture outperforms existing emotion recognition algorithms.

Theoretical Analysis: We mathematically prove that simply increasing the number of model parameters while keeping the dataset size constant does not improve the model performance. Specifically, given a data set \mathcal{D} , $\mathcal{D} = \{(x^1, \hat{y}^1), (x^2, \hat{y}^2), \dots, (x^N, \hat{y}^N)\}$, the actual trained model and the ideal model are defined as follows:

$$\begin{aligned} h^{\text{train}} &= \arg \min_h L(h, \mathcal{D}_{\text{train}}) \\ h^{\text{all}} &= \arg \min_h L(h, \mathcal{D}_{\text{all}}) \end{aligned} \quad (28)$$

where h^{train} is the model trained on the training set $\mathcal{D}_{\text{train}}$, and h^{all} is the model obtained on all existing training datasets \mathcal{D}_{all} .

We hope that $L(h^{\text{train}}, \mathcal{D}_{\text{all}})$ and $L(h^{\text{all}}, \mathcal{D}_{\text{all}})$ should be as close as possible. Namely:

$$\forall h \in \mathcal{H}, \left| L(h, \mathcal{D}_{\text{train}}) - L(h, \mathcal{D}_{\text{all}}) \right| \leq \frac{\delta}{2} \quad (29)$$

where δ is a number close to 0. \mathcal{H} represents a collection of models under different parameter settings.

Due to $h^{\text{train}} = \arg \min_h L(h, \mathcal{D}_{\text{train}})$, we can get $L(h^{\text{train}}, \mathcal{D}_{\text{all}}) \geq L(h^{\text{train}}, \mathcal{D}_{\text{train}})$. Hence we transform Eq. (29) to get the Eq. (30) as follows:

$$\begin{aligned} L(h^{\text{train}}, \mathcal{D}_{\text{all}}) &\leq L(h^{\text{train}}, \mathcal{D}_{\text{train}}) + \frac{\delta}{2} \\ &\leq L(h^{\text{all}}, \mathcal{D}_{\text{train}}) + \frac{\delta}{2} \\ &\leq L(h^{\text{all}}, \mathcal{D}_{\text{all}}) + \frac{\delta}{2} + \frac{\delta}{2} = L(h^{\text{all}}, \mathcal{D}_{\text{all}}) + \delta \end{aligned} \quad (30)$$

According to Eq. (30), we can simplify it to obtain Eq. (31) as follows:

$$L(h^{\text{train}}, \mathcal{D}_{\text{all}}) - L(h^{\text{all}}, \mathcal{D}_{\text{all}}) \leq \delta \quad (31)$$

Under the condition of satisfying Eq. (31), the model can be trained on the training set to obtain an optimal network parameter. Intuitively, when the optimal parameters h^{train} of the model trained on the dataset $\mathcal{D}_{\text{train}}$ are similar to the parameters h^{all} trained on the dataset \mathcal{D}_{all} , we think that h^{train} is the optimal solution trained on the training set.

Since we cannot collect all the existing training data \mathcal{D}_{all} , we can only obtain a subset $\mathcal{D}_{\text{train}}$ by sampling from \mathcal{D}_{all} . We assume that the $\mathcal{D}_{\text{train}}$ obtained by sampling is bad, then there is at least one h to get the following situation:

$$\left| L(h, \mathcal{D}_{\text{train}}) - L(h, \mathcal{D}_{\text{all}}) \right| > \epsilon, \quad \epsilon = \frac{\delta}{2} \quad (32)$$

According to Hoeffding's inequality, we can know the probability of sampling a bad $\mathcal{D}_{\text{train}}$ given model h as follows:

$$P(\mathcal{D}_{\text{train}} \text{ is bad due to } h) \leq 2 \exp(-2N\epsilon^2) \quad (33)$$

where N represents the number of sampling times.

Therefore, the probability that $\mathcal{D}_{\text{train}}$ is bad in all cases of $h \in \mathcal{H}$ is as follows:

$$\begin{aligned} P(\mathcal{D}_{\text{train}} \text{ is bad}) &= \prod_{h \in \mathcal{H}} P(\mathcal{D}_{\text{train}} \text{ is bad due to } h) \\ &\leq \sum_{h \in \mathcal{H}} P(\mathcal{D}_{\text{train}} \text{ is bad due to } h) \\ &\leq \sum_{h \in \mathcal{H}} 2 \exp(-2N\epsilon^2) \\ &= |\mathcal{H}| \cdot 2 \exp(-2N\epsilon^2) \end{aligned} \quad (34)$$

According to Eq. (34), we can know that when the number of samples is fixed, the probability that $\mathcal{D}_{\text{train}}$ is bad is only related to the complexity $|\mathcal{H}|$ of the model. Therefore, we have mathematically proved that when the data set is fixed, the larger the number of parameters of the model, the worse the generalization performance of the model may be. Therefore, we can conclude that the more complex the model is, the better the performance of the model is. Instead, we need to design an effective architecture to learn the distribution law of the data.

6.5. Comparison of modality margin β

Since the modal margin β is a hyperparameter in this paper, we have conducted extensive experiments to verify the effect of different margins β on emotion recognition. As shown in Fig. 4(a), on the IEMOCAP dataset, our model performs best in emotion recognition with a margin $\beta = 0.8$, with a WF1 value of 70.4%. When the margin is too small (e.g., $\beta = 0.5/0.7$), the contrastive learning ability of the model is poor resulting in still large modal gaps. On the contrary, if the margin is too large (e.g., $\beta = 0.9$), complementary semantic information between different modalities may be lost. On the MELD dataset, our model performs best in emotion recognition with a margin $\beta = 0.9$, with a WF1 value of 70.4%. Similar to on the IEMOCAP dataset, too small margins can lead to large modal gaps. Therefore, choosing a good margin has an important impact on the training effect of the model.

6.6. Comparison of contrastive learning methods

To explore the effectiveness of our designed graph contrastive learning mechanism, we compared different contrastive learning methods, i.e., Supervised Contrastive Learning (SCL), Supervised Cluster-level Contrastive Learning (SCCL).

As shown in Fig. 4(b), we use RoBERTa-large as our text encoder to obtain rich context semantic information. For the results on different comparative learning methods, SCL achieves a 1.4% improvement over

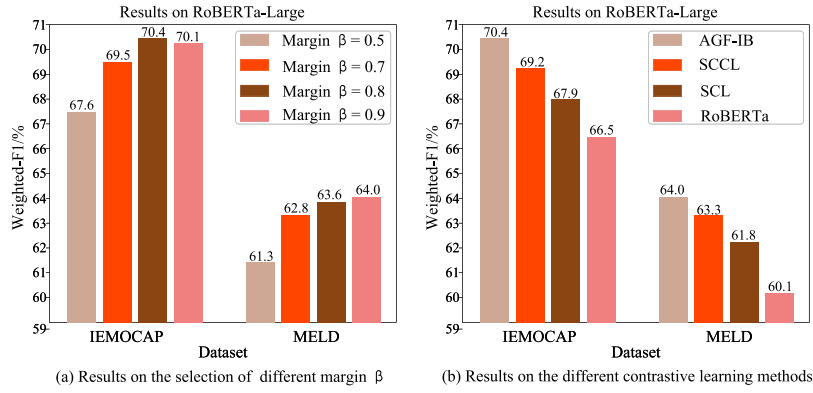


Fig. 4. Experimental results on RoBERTa-Large. (a) Effect of different modal margins β on model training results. (b) Effect of different contrastive learning methods on model training results.

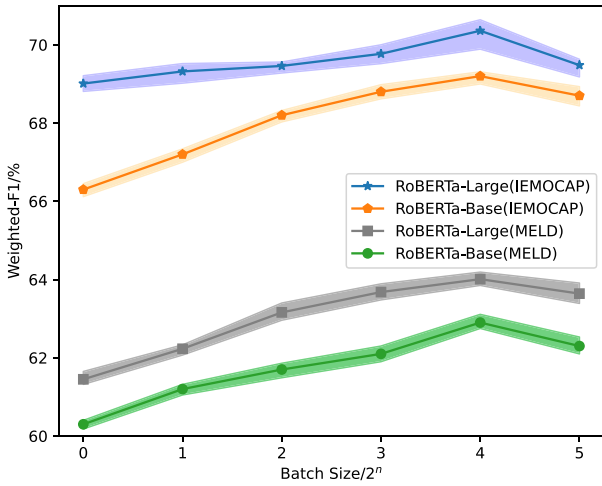


Fig. 5. We use different batch sizes with RoBERTa-Large to verify the stability experiments on IEMOCAP and MELD datasets.

Table 7

Experimental results of C-MFN method on the UR-FUNNY dataset for the humor detection task. C-MFN (C) means using only contextual information without punchlines. C-MFN (P) means using only punchlines with no contextual information. * means the method equipped with the TGAN, IMCL, and ICCL module. The best result is highlighted in bold.

UR-FUNNY					
Modality	T	A+V	T+A	T+V	T+A+V
C-MFN(P)	62.85	53.30	63.28	63.22	64.47
C-MFN(C)	57.96	50.23	57.78	57.99	58.45
C-MFN	64.44	57.99	64.47	64.22	65.23
C-MFN(P)*	67.19	60.87	67.89	68.36	68.97
C-MFN(C)*	61.86	55.36	59.26	60.11	62.49
C-MFN*	68.69	62.06	67.68	67.92	68.28

the RoBERTa baseline on the IEMOCAP dataset and a 1.7% improvement on the MELD dataset. The performance improvement of SCCL on the IEMOCAP and MELD datasets is higher than SCL, achieving 2.7% and 3.2% improvements respectively compared to the RoBERTa baseline. AGF-IB has the highest performance improvement on the IEMOCAP and MELD datasets, achieving 3.9% and 3.9% improvements compared to the RoBERTa baseline, respectively.

The above experimental phenomena illustrate the effectiveness of our designed intra-modal and inter-modal, and intra-class and inter-class contrastive learning mechanism.

6.7. Batch size stability

We use different batch sizes to verify the stability of model training on IEMOCAP and MELD datasets. As shown in Fig. 5, We set the batch size to range from $2^0 = 1$ to $2^5 = 32$. According to the experimental results, the model has the best emotion classification effect when the batch size is 16. When each training step sets a small batch size (i.e., when only a small number of samples are used), the model cannot extract effective features in different modalities, and their contrastive learning effect will be relatively poor.

6.8. Extended research

To verify the scalability of our fusion and contrastive mechanism to other multimodal studies, we apply our method to the task of multimodal humor detection. As shown in Table 7, we utilize Contextual Memory Fusion Network (C-MFN) [38] as the backbone network to verify the effectiveness of our proposed algorithm, where C-MFN (C) means using only contextual information without punchlines. C-MFN (P) means using only punchlines with no contextual information, C represents the context, P represents the punchlines, and C-MFN represents using punchlines and contextual information. We embed our TGAN, IMCL, and ICCL mechanisms into the C-MFN method, and experimental results show that our method outperforms the C-MFN method in any combination of modalities. Specifically, we use accuracy as the evaluation metric for humor detection, and our method can achieve improvements ranging from 1.48% to 7.57%. Experimental results show that our method can be applied not only to multimodal emotion recognition in conversations tasks, but also to other multimodal tasks.

6.9. Ablation study

To verify the rationality of our module design, we use RoBERTa-Base and RoBERTa-Large as our text encoders to conduct ablation experiments. As shown in Fig. 6(b), for the results on RoBERTa-Large, AGF-IB achieves the best experimental results with F1 values of 70.4% and 64.0% on the IEMOCAP and MELD datasets, respectively. The emotional effect of RoBERTa-Large with IMCL is second, and the F1 values are 68.9% and 63.5%, respectively. The emotional effect of RoBERTa-Large with ICCL is worse than RoBERTa-Large with IMCL, and the F1 values are 68.1% and 62.9% respectively. The emotional effect of RoBERTa-Large with TGAN is only slightly better than the RoBERTa-Large baseline, with F1 values of 67.3% and 61.7%, respectively. The experimental results show that the intra-modal and inter-modal contrastive learning is the most critical for the training of the model, which is beneficial for the model to fuse complementary multi-modal semantic information. Intra-class and inter-class contrastive learning is also important for the training of the model,

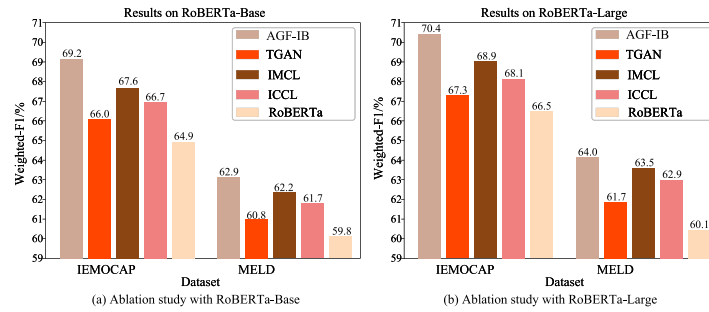


Fig. 6. Ablation experiments on IEMOCAP and MELD datasets. (a) We use RoBERTa-Base as a text encoder to explore the impact of TGAN, IMCL, and ICCL on model training. (b) We use RoBERTa-Large as a text encoder to explore the impact of TGAN, IMCL, and ICCL on model training.

Table 8

The influence of each component on the emotion recognition performance in the IEMOCAP dataset. RoBERTa-Large is chosen as our architecture.

Row Number	Components				IEMOCAP		MELD	
	RoBERTa	TGAN	IMCL	ICCL	WAA	WAF1	WAA	WAF1
1	-	-	-	-	65.21	64.14	58.71	58.03
2	✓	-	-	-	67.21	66.53	60.95	60.17
3	✓	✓	-	-	68.38	67.74	61.93	62.66
4	✓	✓	✓	-	69.52	69.11	63.02	63.87
5	✓	✓	✓	✓	70.46	70.36	64.14	64.01

Table 9

Experimental results with our method and other graph contrastive learning methods on IEMOCAP and MELD datasets. The best result in each column is in bold.

Methods	IEMOCAP				MELD			
	Train		Test		Train		Test	
	WAA	WF1	WAA	WF1	WAA	WF1	WAA	WF1
DGI	88.76	86.54	66.49	63.77	75.11	75.94	61.08	62.23
InfoGraph	90.15	91.37	65.19	64.36	77.85	76.47	60.99	60.37
AGF-IB	88.49	89.05	70.46	70.36	78.94	79.21	64.14	64.01

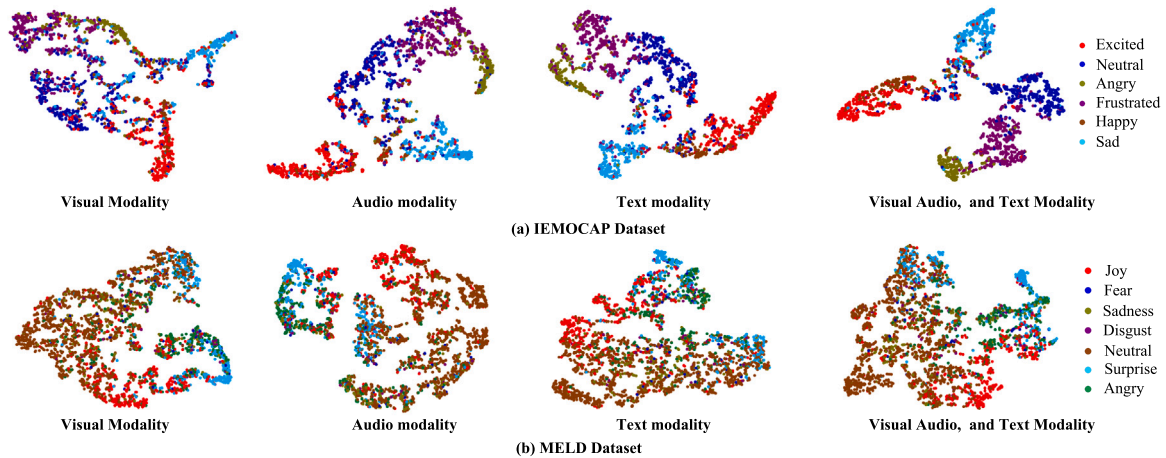


Fig. 7. Visualizing the learned features on the IEMOCAP and MELD benchmark dataset. Each dot represents an utterance, and its color represents an emotion.

which facilitates the class boundary learning of the model. Removing the heterogeneity of modalities is the basis for subsequent model learning.

As shown in Fig. 6(a), for the results on RoBERTa-Base, similar conclusions are drawn from the results of RoBERTa-Base. In addition, the emotion recognition effect of RoBERTa-Large is better than RoBERTa-Base.

Furthermore, we also verified the influence between each component. As shown in Table 8, when no component (only DialogueGCN is used) is used, the accuracy of the model on the IEMOCAP and MELD datasets is 65.21% and 58.71%, and the F1 values are 64.14% and

64.01%, respectively. When only RoBERTa-Large is used, the accuracy of the model on the IEMOCAP and MELD datasets is 67.21% and 60.95%, and the F1 values are 66.53% and 60.17%, respectively. The effect of emotion recognition is better than only using DialogueGCN. When using RoBERTa-Large and TGAN, the emotion recognition effect of the model is better than only using RoBERTa-Large. When using RoBERTa-Large, TGAN and IMCL, the emotion recognition effect of the model is further improved. When four modules of RoBERTa-Large, TGAN, IMCL and ICCL are used, the emotion recognition effect of the model is the best. Experiments demonstrate the effectiveness of each constituent.

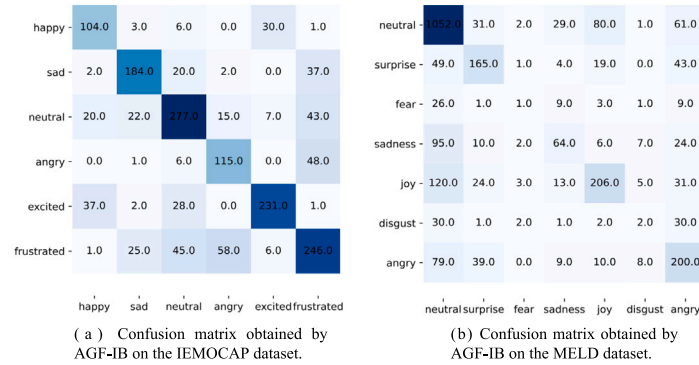


Fig. 8. Confusion matrix of test set on IEMOCAP and MELD datasets. The rows and columns represent the true labels and predicted labels respectively.

Moreover, to illustrate the effectiveness of the proposed graph contrastive learning mechanism that minimizes mutual information, we compare AGF-IB with two existing graph contrastive learning methods that maximize mutual information (i.e., DGI [39] and InfoGraph [40]). Specifically, we replace the proposed graph contrastive learning module with DGI and InfoGraph for experimental verification. The experimental results are shown in Table 9. On the IEMOCAP dataset, although InfoGraph has the best performance on the training set, it has the worst emotion recognition effect on the test set. The emotion recognition effect of DGI on both the training set and the test set is lower than that of the proposed method. On the MELD dataset, the proposed method has the best emotion recognition effect on both the training set and the test set. The experimental results show the effectiveness and robustness of the proposed method.

6.10. Visualization

To display the emotional feature vectors in a high-dimensional space more intuitively, we use the t-SNE method to reduce the dimensionality of the emotional features obtained after model learning in the IEMOCAP and MELD datasets, and obtain a two-dimensional feature embedding. As shown in Fig. 7(a), in the IEMOCAP dataset we can see that the features learned when the model only uses the video modality are relatively confusing, and there are many overlaps between each emotion category. Therefore, it is difficult for the model to correctly classify each emotion category when it only uses video modality. The features learned when the model only uses the audio modality are better than when only the video modality is used, and there are fuzzy class boundaries between different emotion categories. The features learned when the model only uses text modality are better than when only audio or video modality is used. The best features are learned when the model combines text, video and audio modalities, and there are relatively clear class boundaries between different emotion categories, which allows the model to better classify emotions. Experimental results demonstrate the necessity of multi-modal features for the MERC task and the effectiveness of the AGF-IB proposed in this paper. AGF-IB can eliminate the heterogeneity between modalities and effectively capture the intra-modal and inter-modal complementary semantic information. Similarly, as shown in Fig. 7(b), on the MELD dataset, the features learned when the model combines text, video and audio modalities are the best, and there are relatively clear class boundaries between different emotional categories. The experimental results further prove the effectiveness of the AGF-IB proposed in this paper.

6.11. Error analysis

Although the emotion classification effect of the proposed AGF-IB is relatively good, it still cannot correctly classify some minority

categories of emotions and some emotions with similar semantics. Specifically, we analyze the confusion matrix of the test set on the IEMOCAP and MELD datasets. As shown in Fig. 8, AGF-IB cannot classify emotions with similar semantics very well. For example, on the IEMOCAP data set, it is easy to misclassify the “happy” emotion as “excited” and the “angry” emotion as “frustrated”. It is easy to classify the “surprise” sentiment into “angry” on the MELD dataset. AGF-IB also has a classification preference for “neutral” emotions, because the MELD dataset has data imbalance problems, and “neutral” emotions belong to most categories. It is difficult for the model to classify the two emotions of “fear” and “disgust” on MELD because the number of samples for these two emotions is very small.

To demonstrate the efficacy of the proposed AGF-IB, we tested some cases. As shown in Fig. 9, we selected a conversation on the MELD dataset. AGF-IB, which uses multi-modal features, can correctly classify the emotions of all utterances, while DialogueRNN and DialogueGCN incorrectly predict the third utterance as “surprise” and the fourth and fifth utterances as “neutral”. Experimental results show that DialogueRNN and DialogueGCN cannot make good use of complementary semantic information within and between modalities, while AGF-IB has more powerful multi-modal fusion capabilities. On the other hand, using only the text modality, AGF-IB incorrectly identifies the fourth and fifth utterances as “neutral”. The experimental results further illustrate the necessity of multi-modal features for the MERC task.

7. Conclusion and future work

In this paper, we propose a novel Adversarial Alignment and Graph Fusion via Information Bottleneck for the Multimodal Emotion Recognition in Conversations architecture (AGF-IB) model, which enables cross-modal feature fusion, intra-modal and inter-modal contrasting representation learning, and intra-class and inter-class representation learning. In addition, AGF-IB uses information bottlenecks to minimize the mutual information between multiple views to obtain structurally heterogeneous but semantically similar multiple views. Specifically, we firstly introduce a cross-modal feature fusion method based on adversarial learning to eliminate the heterogeneity among different modalities. Secondly, to comprehensively consider the relationship between intra-modality and inter-modality and the relationship between intra-class and inter-class, and obtain a compact node representation, we design a novel graph contrastive learning architecture via IB to enhance the representation ability of nodes by increasing the distance between different emotion labels of the same modality and shrinking the distance between the same emotion of different modalities, and minimizing MI between views. Finally, we use a multi-layer perceptron (MLP) for emotion classification.

In future work, we consider using diffusion models for feature fusion across modalities to generate fused features that contain more semantic information. In addition, we will also consider transferring our method to other multimodal tasks.











Turn	Speaker	Visual	Audio	Text	Dialogue RNN	Dialogue GCN	Ours (only text)	Ours	Ground Truth
1	Joey			Oh my god, you're back!	surprise	surprise	surprise	surprise	surprise
2	Phoebe			Ohh, let me see it! Let me see your hand!	surprise	surprise	surprise	surprise	surprise
3	Monica			Why do you want to see my hand?	<i>surprise</i>	<i>surprise</i>	neutral	neutral	neutral
4	Phoebe			I wanna see what's in your hand. I wanna see the trash.	<i>neutral</i>	<i>neutral</i>	<i>neutral</i>	disgust	disgust
5	Phoebe			Eww! Oh, it's all dirty. You should throw this out.	<i>neutral</i>	<i>neutral</i>	<i>neutral</i>	disgust	disgust

Fig. 9. In the MELD dataset, we selected a conversation to test the emotion recognition performance of DialogueRNN, DialogueGCN and AGF-IB.

CRedit authorship contribution statement

Yuntao Shou: Conceptualization, Formal analysis, Methodology, Software, Writing – original draft. **Tao Meng:** Funding acquisition, Investigation, Methodology, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. **Wei Ai:** Funding acquisition, Supervision, Writing – review & editing. **Fuchen Zhang:** Investigation, Software, Writing – review & editing. **Nan Yin:** Software, Writing – review & editing. **Keqin Li:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors deepest gratitude goes to the anonymous reviewers and AE for their careful work and thoughtful suggestions that have helped improve this paper substantially. This work is supported by National Natural Science Foundation of China (Grant No. 69189338, Grant No. 62372478), Excellent Young Scholars of Hunan Province of China (Grant No. 22B0275), Changsha Natural Science Foundation (Grant No. kq2202294), and program of Research on Local Community Structure Detection Algorithms in Complex Networks (Grant No. 2020YJ009).

References

- [1] F. Huang, X. Li, C. Yuan, S. Zhang, J. Zhang, S. Qiao, Attention-emotion-enhanced convolutional LSTM for sentiment analysis, *IEEE Trans. Neural Netw. Learn. Syst.* 33 (9) (2022) 4332–4345.
- [2] S.K. Khare, V. Bajaj, Time–frequency representation and convolutional neural network-based emotion recognition, *IEEE Trans. Neural Netw. Learn. Syst.* 32 (7) (2020) 2901–2909.
- [3] S. Qian, D. Xue, Q. Fang, C. Xu, Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval, *IEEE Trans. Pattern Anal. Mach. Intell.* (2022) 1–18.
- [4] J. Guo, B. Song, P. Zhang, M. Ma, W. Luo, et al., Affective video content analysis based on multimodal data fusion in heterogeneous networks, *Inf. Fusion* 51 (2019) 224–232.
- [5] W. Zhao, Y. Zhao, X. Lu, Cauain: Causal aware interaction network for emotion recognition in conversations, in: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI, Morgan Kaufmann, 2022*, pp. 4524–4530.
- [6] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017*, pp. 1103–1114.
- [7] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A.B. Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018*, pp. 2247–2256.
- [8] J. Hu, Y. Liu, J. Zhao, Q. Jin, MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021*, pp. 5666–5675.
- [9] S. Liu, P. Gao, Y. Li, W. Fu, W. Ding, Multi-modal fusion network with complementarity and importance for emotion recognition, *Inform. Sci.* 619 (2023) 679–694.
- [10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.
- [11] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: An attentive rnn for emotion detection in conversations, in: *Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, (01) 2019*, pp. 6818–6825.
- [12] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: *Proceedings of the 18th ACM International Conference on Multimedia, 2010*, pp. 1459–1462.
- [13] Z. Lian, B. Liu, J. Tao, PIRNet: Personality-enhanced iterative refinement network for emotion recognition in conversation, *IEEE Trans. Neural Netw. Learn. Syst.* (2022) 1–12.
- [14] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL, 2017*, pp. 873–883.
- [15] R. Beard, R. Das, R.W.M. Ng, P.G.K. Gopalakrishnan, L. Eerens, P. Swietojanski, O. Miksik, Multi-modal sequence fusion via recursive attention for emotion recognition, in: *Proceedings of the 22nd Conference on Computational Natural Language Learning, ACL, 2018*, pp. 251–259.
- [16] M. Ren, X. Huang, W. Li, D. Song, W. Nie, LR-GCN: Latent relation-aware graph convolutional network for conversational emotion recognition, *IEEE Trans. Multimed.* (2021) 1.
- [17] W. Nie, M. Ren, J. Nie, S. Zhao, C-GCN: correlation based graph convolutional network for audio-video emotion recognition, *IEEE Trans. Multimed.* 23 (2020) 3793–3804.
- [18] S. Wu, L. Zhou, Z. Hu, J. Liu, Hierarchical context-based emotion recognition with scene graphs, *IEEE Trans. Neural Netw. Learn. Syst.* (2022) 1–15.
- [19] C.-M. Chang, C.-C. Lee, Learning enhanced acoustic latent representation for small scale affective corpus with adversarial cross corpora integration, *IEEE Trans. Affect. Comput.* (2021) 1.
- [20] M. Li, B. Yang, J. Levy, A. Stolcke, V. Rozgic, S. Matsoukas, C. Papayiannis, D. Bone, C. Wang, Contrastive unsupervised learning for speech emotion recognition, in: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021*, pp. 6329–6333.
- [21] D. Kim, B.C. Song, Contrastive adversarial learning for person independent facial emotion recognition, in: *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, (7) AAAI, 2021*, pp. 5948–5956.
- [22] X. Wang, D. Zhang, H.-Z. Tan, D.-J. Lee, A self-fusion network based on contrastive learning for group emotion recognition, *IEEE Trans. Comput. Soc. Syst.* (2022) 1–12.
- [23] T. Kim, P. Vossen, Emoberta: Speaker-aware emotion recognition in conversation with Roberta, *Comput. Res. Repos.-arXiv 2021 (2021) 1–7*.

- [24] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, N. Onoe, M2fnet: Multi-modal fusion network for emotion recognition in conversation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4652–4661.
- [25] W. Shen, S. Wu, Y. Yang, X. Quan, Directed acyclic graph network for conversational emotion recognition, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 1551–1560.
- [26] Z. Li, F. Tang, M. Zhao, Y. Zhu, EmoCaps: Emotion capsule based model for conversational emotion recognition, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 1610–1618.
- [27] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, DialogueGCN: A graph convolutional neural network for emotion recognition in conversation, 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, ACL, 2019, pp. 154–164.
- [28] A.A. Alemi, I. Fischer, J.V. Dillon, K. Murphy, Deep variational information bottleneck, in: International Conference on Learning Representations, 2016.
- [29] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: A multimodal multi-party dataset for emotion recognition in conversations, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL, 2019, pp. 527–536.
- [30] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, *Lang. Resour. Eval.* 42 (4) (2008) 335–359.
- [31] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2015.
- [32] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, ACL, 2014, pp. 1746–1751.
- [33] Z. Lian, B. Liu, J. Tao, CTNet: Conversational transformer network for emotion recognition, *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 29 (2021) 985–1000.
- [34] D. Hu, X. Hou, L. Wei, L. Jiang, Y. Mo, MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 7037–7041.
- [35] W. Han, H. Chen, S. Poria, Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 9180–9192.
- [36] S. Mai, H. Hu, 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, (01) 2020, pp. 164–172.
- [37] Z. Lian, L. Chen, L. Sun, B. Liu, J. Tao, Gcnet: Graph completion network for incomplete multimodal learning in conversation, *IEEE Trans. Pattern Anal. Mach. Intell.* (2023).
- [38] M.K. Hasan, W. Rahman, A.B. Zadeh, J. Zhong, M.I. Tanveer, L.-P. Morency, M.E. Hoque, UR-FUNNY: A multimodal language dataset for understanding humor, 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, 2019, pp. 2046–2056.
- [39] P. Veličković, W. Fedus, W.L. Hamilton, P. Liò, Y. Bengio, R.D. Hjelm, Deep graph infomax, in: International Conference on Learning Representations, 2018.
- [40] F.-Y. Sun, J. Hoffman, V. Verma, J. Tang, InfoGraph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization, in: International Conference on Learning Representations, 2019.