DER-GCN: Dialog and Event Relation-Aware Graph Convolutional Neural Network for Multimodal Dialog Emotion Recognition

Wei Ai, Yuntao Shou, Tao Meng, and Keqin Li, *Fellow, IEEE*

Abstract*— With the continuous development of deep learning (DL), the task of multimodal dialog emotion recognition (MDER) has recently received extensive research attention, which is also an essential branch of DL. The MDER aims to identify the emotional information contained in different modalities, e.g., text, video, and audio, and in different dialog scenes. However, the existing research has focused on modeling contextual semantic information and dialog relations between speakers while ignoring the impact of event relations on emotion. To tackle the above issues, we propose a novel dialog and event relation-aware graph convolutional neural network (DER-GCN) for multimodal emotion recognition method. It models dialog relations between speakers and captures latent event relations information. Specifically, we construct a weighted multirelationship graph to simultaneously capture the dependencies between speakers and event relations in a dialog. Moreover, we also introduce a self-supervised masked graph autoencoder (SMGAE) to improve the fusion representation ability of features and structures. Next, we design a new multiple information Transformer (MIT) to capture the correlation between different relations, which can provide a better fuse of the multivariate information between relations. Finally, we propose a loss optimization strategy based on contrastive learning to enhance the representation learning ability of minority class features. We conduct extensive experiments on the benchmark datasets, Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Multimodal EmotionLines Dataset (MELD), which verify the effectiveness of the DER-GCN model. The results demonstrate that our model significantly improves both the average accuracy and the $F1$ value of emotion recognition. Our code is publicly available at https://github.com/yuntaoshou/DER-GCN.

***Index Terms*— Contrastive learning, event extraction, masked graph autoencoders (MGAEs), multimodal dialog emotion recognition (MDER), multiple information Transformer (MIT).**

Manuscript received 1 December 2022; revised 21 June 2023, 23 October 2023, and 4 January 2024; accepted 15 February 2024. Date of publication 4 March 2024; date of current version 1 March 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 69189338, in part by the Excellent Young Scholars of Hunan Province of China under Grant 22B0275, and in part by Changsha Natural Science Foundation under Grant kq2202294. *(Corresponding author: Tao Meng.)*

Wei Ai, Yuntao Shou, and Tao Meng are with the School of Computer and Information Engineering, Central South University of Forestry and Technology, Changsha, Hunan 410004, China (e-mail: aiwei@hnu.edu.cn; yuntaoshou@csuft.edu.cn; mengtao@hnu.edu.cn).

Keqin Li is with the Department of Computer Science, State University of New York, New Paltz, NY 12561 USA (e-mail: lik@newpaltz.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2024.3367940, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2024.3367940

I. Introduction

A. Motivation

THE task of multimodal dialog emotion recognition (MDER) is to identify the emotional changes of speakers in different modalities, such as text, video, and audio. In recent decades, due to the application (APP) of MDER in some emerging APP scenarios, for instance, the recognition of negative emotions has attracted research attention in social media, such as Meta and Weibo [1], the intelligent recommendation system for online shopping [2] and chat robots [3]. Furthermore, when shopping online, the APP will recommend the most interesting products according to the user's preferences.

However, MDER is more challenging than sentence-level emotion recognition or unimodal emotion recognition tasks, because sentiment changes are generally determined by a series of meaningful internal and external factors. Specifically, in the dialog process, the speaker's emotion is affected not only by internal factors composed of contextual information but also by external factors composed of dialog and event relationships (e.g., entity, location, keywords, and so on). Details of dialog and event relationships are provided in the Supplementary Material. For example, when speakers talk about a sensitive topic on social media, they often express their emotions more implicitly and suggestively. Therefore, events can be exploited to strengthen conversational semantic relationships between speakers, thereby compensating for the lack of explicit semantic features. However, how to comprehensively consider the influence of internal and external factors on emotion recognition is still a problem to be solved. In addition, in MDER, due to the high cost of labeling, the data distribution exhibits a long-tailed state. It leads to the model being less effective at identifying the minority class emotion.

The current mainstream MDER methods utilize recurrent neural networks (RNNs) [4], Transformers [5], and graph neural networks (GNNs) [6] to model the semantic information of context and dialog relationship between speakers, respectively. To better integrate contextual semantic information, Transformer-based methods are applied, but they still ignore the influence of external factors on emotion recognition. To consider the influence of internal and external factors on emotion recognition, many researchers have begun to adopt GNN to model MDER. Although the abovementioned methods have achieved good results in emotion recognition, they all

2162-237X © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 03,2025 at 01:07:51 UTC from IEEE Xplore. Restrictions apply.

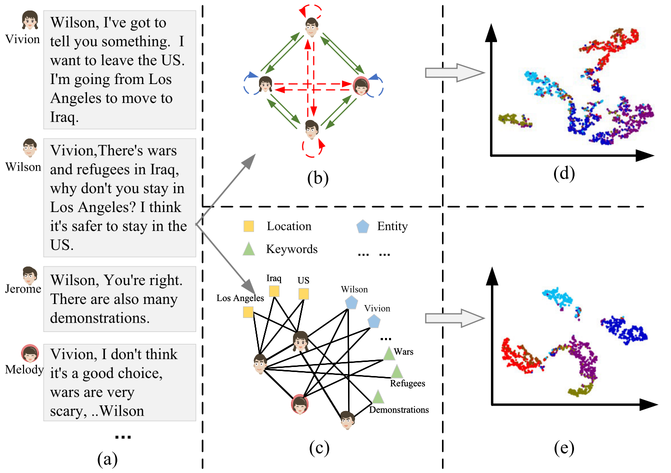

Vivion
Wilson, I've got to tell you something. I want to leave the US. I'm going from Los Angeles to move to Iraq.
Wilson
Vivion,There's wars and refugees in Iraq, why don't you stay in Los Angeles? I think it's safer to stay in the US.
Jerome
Wilson, You're right. There are also many demonstrations.
Melody
Vivion, I don't think it's a good choice, wars are very scary, ..Wilson
...
(a)
(b)
(c)
(d)
(e)
Location
Entity
Keywords
... ...
Los Angeles
Iraq
US
Wilson
Vivion
...
Wars
Refugees
Demonstrations

Fig. 1. Illustrative example of the impact of event relationships on the feature embeddings of sentences. (a) Raw dialog text with four speakers. (b) Graph of dialog relationships composed of emotional interactions between speakers. (c) Emotional interaction graph is composed of dialog and event relationships. (d) Feature embeddings of sentences in graphs composed of dialog relations. (e) Feature embeddings of sentences in graphs composed of dialog and event relationships. In (d) and (e), points of different colors represent different emotion categories.

ignore the influence of external factors (i.e., event relationships). However, the event relationship also greatly influences the speaker's emotion, and the speaker usually shows the same emotion when discussing the same event. Therefore, modeling the event relations in the dialog is beneficial to obtain a better feature embedding of emotion categories. As shown, Fig. 1(b) is a graph that only considers the dialog relationship between speakers. Its emotion categories have many overlapping areas in the feature embeddings, and there are no clear class boundaries between each emotional category, as shown in Fig. 1(d). Fig. 1(c) is a graph that comprehensively considers the interaction relationship and event relationship between speakers. Its feature embedding of emotion categories has better discrimination, as shown in Fig. 1(e). Hence, it is necessary to take the event relationship as the starting point of the MDER architecture design.

To tackle the above problem, we propose a novel dialog and event relation-aware graph convolutional neural network (DER-GCN) for multimodal emotion recognition architecture. DER-GCN mainly includes six modules: data preprocessing, feature extraction and fusion, mask graph representation learning, multirelational information aggregation, balanced sampling strategy, and emotion classification. First, we use robustly optimized BERT approach (RoBERTa) [7], 3D convolutional neural network (3D-CNN) [8], and bidirectional long short-term memory (Bi-LSTM)-based encoder [9] to obtain embedding representations for three modalities: text, video, and audio. Second, we use a bidirectional gated recurrent unit (Bi-GRU) for feature extraction and Doc2EDAG [10] for event extraction to strengthen the dialog relationship between speakers. Then, we design a novel cross-modal feature fusion method to learn complementary semantic information between different modalities. Specifically, we use cross-modal attention to learn the differences between the semantic information of different modes. The average pooling operation is used to learn the global information of each mode to guide the intermodality and intramodality information aggregation, respectively. Third, we design a self-supervised mask graph autoencoder (SMGAE) to model the correlation between dialogs and events. Unlike the previous works [11], which only perform mask reconstruction on nodes in the graph, SMGAE performs mask reconstruction on some nodes and edges simultaneously. Fourth, we design the multiple information Transformer (MIT) to better fuse the multivariate information between relations and capture the correlation between different relations. MIT is a paid attention mechanism to filter unimportant relational information, which fuses to obtain better embedding representations. Fifth, we propose a loss optimization function based on contrastive learning to alleviate the long-tailed effect in MDER, which balances the proportion of each emotion category during model training. Finally, we have used an emotion classifier constructed from a multilayer perceptron (MLP) to output the final sentiment category.

B. Our Contributions

Hence, MDER should consider the dialog between speakers and the event relationship in the dialog as the starting point of model design. Inspired by the analysis above, we present a novel DER-GCN for multimodal emotion recognition to learn better emotion feature embedding. The main contributions of this article are summarized as follows.

1) A novel dialog and event relation-aware emotion representation learning architecture is presented and named DER-GCN. DER-GCN can achieve cross-modal feature fusion, solve the imbalanced data distribution problem, and learn more discriminative emotion class boundaries.
2) A novel self-supervised graph representation learning framework, named SMGAE, is presented. SMGAE enhances the feature representation capability of nodes and optimizes the structural representation of graphs, which has a stronger antinoise ability.
3) A new weighted relation-aware multiple subgraph information aggregation method is implemented and named MIT. MIT is used to learn the importance of different relations in information aggregation to fuse to obtain more discriminative feature embeddings.
4) Finally, extensive experiments are performed on two popular benchmark datasets, Multimodal EmotionLines Dataset (MELD) and Interactive Emotional Dyadic Motion Capture (IEMOCAP), which demonstrate that DER-GCN outperforms the existing comparative algorithms in weight accuracy (WA) and $F1$ value for multimodal emotion recognition.

II. Related Work

A. Emotion Recognition in Conversation

MDER is an interdisciplinary research field that has attracted extensive attention from researchers in cognitive science, psychology, and so on. The existing MDER research mainly includes emotion recognition based on RNN [12], emotion recognition based on GNN [13], and emotion recognition based on Transformer [5]. RNNs mainly extract contextual semantic information by modeling long-range contextual dependencies. The GNNs model, the dynamic interaction process of dialog, mainly relies on the graph structure's inherent properties to model the dependencies between speakers. The Transformer mainly uses the attention mechanism to achieve

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 03,2025 at 01:07:51 UTC from IEEE Xplore. Restrictions apply.

cross-modal feature fusion to capture the different semantic information between modalities.

In the RNN-based multimodal emotion recognition research, Wang et al. [14] conducted dual-sequence LSTM (DS-LSTM), which uses a dual-stream LSTM to extract contextual features in the Mei-Frequency map simultaneously. DS-LSTM comprehensively considers the context features of different times and frequencies and achieves a better emotion recognition effect. Li et al. [15] created attention-based bidirectional LSTM RNNs (A-BiLSTM RNNs). This method combines the self-attention mechanism and LSTM to learn multimodal features with a time dimension. Although RNN-based methods have achieved good results in emotion recognition tasks based on contextual semantic modeling, they still ignore the influence of external factors (e.g., dialog relations and event relations).

In Transformer-based multimodal emotion recognition research, Huang et al. [16] employed multimodal Transformer fusion (MTF), which uses a multihead attention mechanism to obtain intermediate feature representations of multimodal emotions. Then, a self-attention mechanism is utilized to capture long-lived dependencies in context. Transformer-based methods can extract richer contextual semantic information, but they still ignore the influence of external factors on emotion recognition.

In GNN-based multimodal emotion recognition research, Sheng et al. [17] performed a summarization and aggregation graph inference network (SumAggGIN), which captures distinguishable fine-grained features between phrases by building a heterogeneous GNN. Although the GNN-based method considers the dialog relationship, it still ignores the influence of the event relationship on MDER.

B. Transformers for Dialog Generation

In recent years, the task of dialog generation has also begun to receive extensive attention. Huang et al. [18] proposed persona-adaptive attention (PAA), which uses a dynamic mask attention mechanism to adaptively reduce redundant information in context information. For example, dialog generation technology can be used in healthcare to help patients access health information. Zheng et al. [19] proposed a pretrained personalized dialog model, which uses a large-scale pretrained model to initialize model weights and introduces attention in the decoder to dynamically extract context information and role information. Zeng and Nie [20] introduced a condition-aware Transformer to generate probability deviations for words in different positions.

C. Masked Self-Supervised Graph Learning

Masked self-supervised graph representation learning, which can automatically learn deeper feature representations from raw data without using a large amount of labeled data, has been used by more and more researchers. The current mainstream research on mask self-supervised graph representation learning focuses on mask and data reconstruction at the node and edge levels.

In node-level mask-based self-supervised learning, Liu et al. [21] performed a spatiotemporal graph neural network (STG-Net), which masks graph nodes based on an edge weighting strategy. GCN is used to reconstruct contextual features to obtain a better data representation. Wang et al. [23] created HeCo, which learns high-level embedding representations of nodes by using a view masking mechanism. In addition, HeCo introduces a contrastive learning strategy, which can further improve the model's ability to learn feature representations.

In edge-level mask-based self-supervised learning, Pan et al. [22] conducted adversarial graph embedding (AGE), which reconstructs the topology of a graph by using an adversarial regularized graph autoencoder (ARGA) and an adversarial regularized variational graph autoencoder (ARVGA). AGE is trained in a self-supervised manner to learn the underlying distribution law of the data. The above methods only consider structure mask reconstruction and ignore feature mask reconstruction.

D. Balanced Optimization Based on Contrastive Learning

The datasets in MDER suffer from data imbalance, which makes the cross-entropy loss function widely used for classification no longer applicable. However, contrastive learning can learn distinguishable class boundary information between different classes by continuously narrowing the gap between positive samples. It continuously widens the gap between positive and negative samples [24]. Therefore, contrastive learning is often used to solve the data imbalance problem in practical problems.

Cai et al. [25] applied a heterogeneous graph contrastive learning (HGCL) network, which obtains the embedded representation of each node by maximizing the interaction information between local graph nodes and the global representation of the full graph nodes. HGCL can learn better class boundary information from multivariate heterogeneous data. Peng et al. [26] proposed supervised contrastive learning (SCL) to compare the input samples with other instances and input samples with negative samples, which were generated by the soft Brownian offset sampling method to enhance feature representation capability. SCL can effectively alleviate the problem of imbalanced data distribution by continuously expanding the difference between positive and negative samples.

III. Preliminary Information

In this section, we will define the multimodal emotion recognition task and briefly introduce the preprocessing methods for the three modalities of text, audio, and video in the multimodal emotion dataset. Their processing procedures are as follows: 1) *word embedding*—to obtain word vectors with rich semantic information, we will use the RoBERTa model [7] to obtain the vector representation of each word; 2) *visual feature extraction*—to capture the features of the speaker's facial expression changes and gesture changes in each frame of the video, we will use the 3D-CNN model [27] for feature extraction; and 3) *audio feature extraction*—to capture the speech features that can distinguish different speakers, we would use the structure of the encoder to extract the feature of the sound signal. In addition, to capture the semantic

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 03,2025 at 01:07:51 UTC from IEEE Xplore. Restrictions apply.

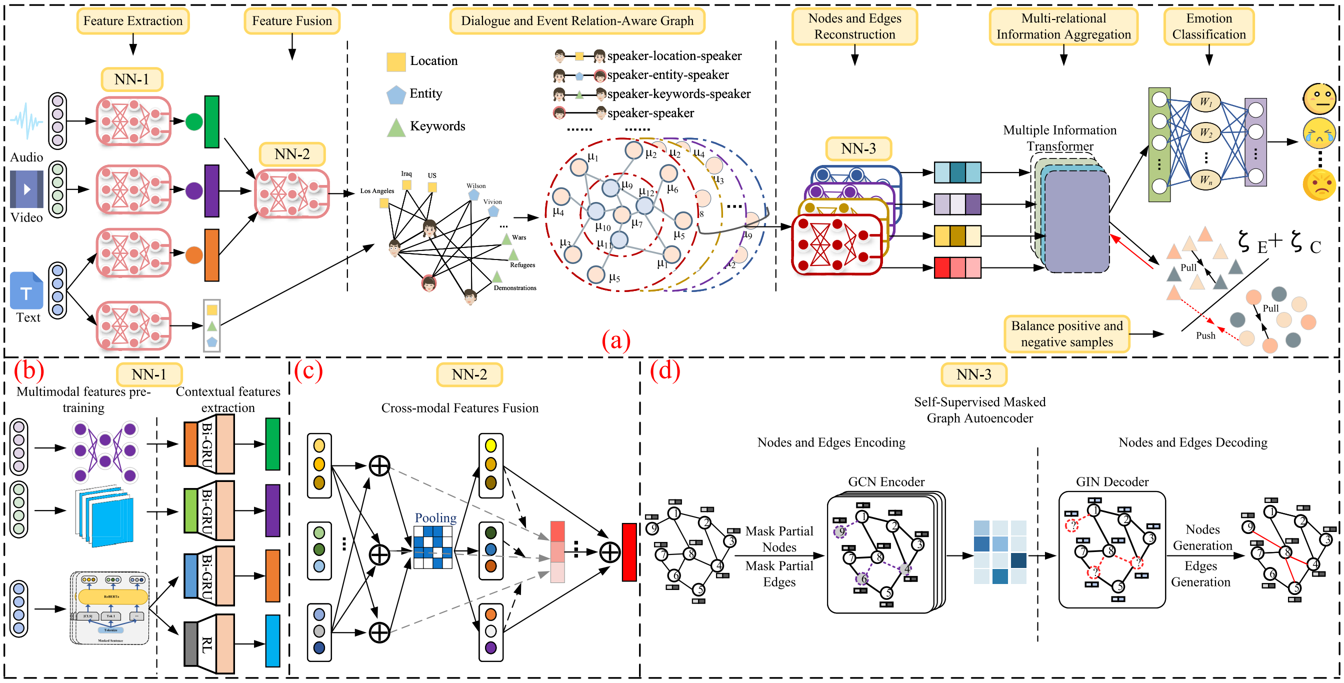


Fig. 2. (a) Overall process framework of DER-GCN: it first preprocesses multimodal data to obtain encoded feature embeddings via NN-1. Second, it uses NN-2 to achieve cross-modal feature fusion. Third, it constructs a weighted multirelational dialog and event relation-aware graph through the fused feature vectors. Fourth, node and edge features are reconstructed via NN-3. Fifth, the fused multirelational information feature vectors are obtained through the MIT, and a loss optimization strategy based on contrastive learning is used to solve the data imbalance problem. Finally, it uses the emotion classifier to get the final emotion label. (b) NN-1: multimodal feature encoder. (c) NN-2: cross-modal feature aggregator. (d) NN-3: self-supervised masked graph encoder.

information of the topic events discussed by the speaker during the dialog, we also perform event extraction on the text.

A. Problem Definition

For the task of MDER, since the number of speakers $N (N \geq 2)$ participating in the dialog is not fixed, we assume that N speakers are involved in a conversation and are represented as $P = \{P_1, P_2, \ldots, P_N\}$, respectively. During the dialog, a series of utterances from the speaker are arranged in a chronological order, which can be expressed as $U = \{u_1, u_2, \ldots, u_T\}$, where T represents the total number of utterances, and each has three modalities, i.e., text (t), audio (a), and visual (v). The task of this article is to predict the speaker's emotion category at the current moment q based on the speaker's words, voice, and expressions. The emotion prediction task is defined as follows:

$$e_q = \text{prediction}\left(\{u_{q-K}, \ldots, u_{q-1}\}\right) \tag{1}$$

where e_q represents the emotion of the qth utterance and K represents the window size of the historical context. In this article, we set $K = 10$.

B. Multimodal Feature Extraction

The input of the MDER task is three modal features: text u_t, audio u_a, and video u_v. For each utterance, we extract text, audio, and video features. The specific extraction process is as follows: 1) *word embedding*—to obtain word vectors with rich semantic information, we will use the RoBERTa model [7] to obtain the vector representation of each word; 2) *audio feature extraction*—to capture the speech features that can distinguish different speakers, we would use the structure of the encoder to extract the feature of the sound signal; and 3) *visual feature extraction*—to capture the features of the speaker's facial expression changes and gesture changes in each frame of the video, we use the 3D-CNN model [8] for feature extraction. In addition, to capture the semantic information of the topic events discussed by the speaker during the dialog, we also perform event extraction on the text.

IV. Methodology

A. Design of the DER-GCN Structure

In this section, we illustrate the six components that make up DER-GCN, as shown in Fig. 2. The structure of DER-GCN is as follows.

1) *Sequence Modeling and Cross-Modal Feature Fusion:* For the input text, video, and audio modal features, DER-GCN inputs them into the Bi-GRU to extract contextual semantic information. Furthermore, to capture the regions with the strongest emotional features among the three modalities, we design a cross-modal attention mechanism for feature extraction and fusion of complementary semantic information.
2) *Multirelational Emotional Interaction Graph:* Unlike the current mainstream algorithms that ignore the impact of event relationships on emotional boundary learning, we construct a multirelational GNN that includes events and speakers, thereby enhancing the feature representation capability of the model.
3) *Intrarelational Masked Graph Autoencoder (MGAE):* To improve the fusion representation ability of node features and edge structures in GCN, we designed an MGAE. MGAE improves the representation ability of

GCN by random masking and reconstruction of nodes and edges and alleviates the problem of class distribution imbalance.

4) *Information Aggregation Between Relations:* To guide DER-GCN better to perform information aggregation of multirelational GNNs, we design a multirelational information fusion Transformer, which can effectively fuse the semantic information in the subgraphs composed of different relationships and learn better-embedded representation.
5) *Contrastive Learning:* The commonly used benchmark datasets in the field of multimodal emotion recognition have the problem of unbalanced class distribution. We introduce a contrastive learning mechanism to learn more discriminative class boundary information.
6) *Emotion Classifier:* To make DER-GCN provide more gradient information in the backpropagation process and promote the model to be fully trained during emotion classification, we construct a linear layer with residual connections as the emotional classifier of DER-GCN.

1) Sequence Modeling and Cross-Modal Feature Fusion: The emotional change of the speaker at the current time t is not only related to the utterance at the tth time but also to the contextual utterances before the $t-1$ time and after the $t+1$ time. However, capturing the contextual semantic information contained in the three modalities of video, audio, and text is a challenging task. In this article, we use Bi-GRU to model the long-term dependencies of the three modalities, so that the model can more accurately understand the emotional changes of the speaker at the current moment t. The formula for GRU is defined as follows:

$$\begin{aligned} z_t^\gamma &= \operatorname{sigmoid}\left(W_z^\gamma \cdot \left[h_{t-1}^\gamma, u_t^\gamma\right]\right) \\ r_t^\gamma &= \operatorname{sigmoid}\left(W_r^\gamma \cdot \left[h_{t-1}^\gamma, u_t^\gamma\right]\right) \\ \tilde{h}_t^\gamma &= \tanh\left(W_{\tilde{h}_t}^\gamma \cdot \left[r_t^\gamma \odot h_{t-1}^\gamma\right]\right) \\ h_t^\gamma &= \left(1 - z_t^\gamma\right) \odot h_{t-1}^\gamma + z_t^\gamma \odot \tilde{h}_t^\gamma \end{aligned} \tag{2}$$

where z_t represents the update gate, which is used to select the context information that needs to be retained at the current time t to update the state of the hidden layer at the $t-1$th time. r_t represents the reset gate, which is used to forget the unimportant contextual information in the conversation at the current moment t. u_t and h_t represent the input unimodal feature vectors and the hidden layer for storing contextual information, respectively. $\tilde{h}_t$ represents the candidate's hidden layer state. W_z, W_r, and $W_{\tilde{h}_t}$ are parameters that can be learned in GRU. $\gamma \in \{t, v, a\}$ represents text, video, and audio, respectively. $\odot$ means Hadamard product.

Bi-GRU contains contextual semantic information extracted from forward and reverse. The formula is defined as follows:

$$\begin{aligned} \delta_t^\gamma &= \left[\overrightarrow{h}_t^\gamma : \overleftarrow{h}_t^\gamma\right] \\ \psi^\gamma &= \operatorname{concat}\left(\left[\delta_1^\gamma, \delta_2^\gamma, \ldots, \delta_T^\gamma\right]\right) \end{aligned} \tag{3}$$

where $\overrightarrow{h}_t^\gamma$ is the contextual semantic information extracted in the forward direction, $\overleftarrow{h}_t^\gamma$ is the contextual information extracted in the reverse direction, δ_t represents the sequential context information extracted by the forward and reverse GRUs at the tth moment, and ψ^γ is composed of all the contextual information at the previous T moments.

To realize the information interaction and fusion among the three modalities, we propose a cross-modal attention mechanism, which is used to exploit the interaction between modalities in a more fine-grained manner to improve the semantic understanding ability of the model.

First, we normalize the hidden layer feature vectors of the three modalities obtained after Bi-GRU processing. The formula is defined as follows:

$$\mathcal{H}_{ij}^\gamma = \frac{\exp\left(\varepsilon^\gamma \psi_{ij}^\gamma\right)}{\sum_{i=1}^n \exp\left(\varepsilon^\gamma \psi_{ij}^\gamma\right)} \tag{4}$$

where $\varepsilon^\gamma = (1/\sqrt{d^\gamma})$ is the scaling factor of the three modalities. n represents the dimension of the modality, and $\mathcal{H}_{ij}$ represents the feature vector of the ith row and jth column.

Then, to better preserve the semantic information of the three modalities, we perform an average pooling operation on H_{ij}^γ, and the formula is defined as follows:

$$\xi_t, \xi_a, \xi_v = f_{\text{pooling}}\left(\mathcal{H}^t, \mathcal{H}^a, \mathcal{H}^v\right) \tag{5}$$

where $f_{\text{pooling}}(\cdot)$ represents the average pooling operation.

Next, we perform a fusion operation on the three modal features and use the tanh activation function to obtain their weights. The formula is defined as follows:

$$\begin{aligned} \omega^t &= W^t \tanh\left(\lambda_v \mathcal{H}^v + \lambda_a \mathcal{H}^a + \lambda_t \xi_t + b_t\right) \\ \omega^a &= W^a \tanh\left(\lambda_t \mathcal{H}^t + \lambda_v \mathcal{H}^v + \lambda_a \xi_a + b_a\right) \\ \omega^v &= W^v \tanh\left(\lambda_t \mathcal{H}^t + \lambda_a \mathcal{H}^a + \lambda_v \xi_v + b_v\right) \end{aligned} \tag{6}$$

where W^t, W^a, W^v, λ_a, λ_v, λ_t, b_t, b_a, and b_v are the network parameters that can be learned in the model. According to the above formula, we can get the normalized attention weight $\widetilde{\omega}^\gamma \in \{\widetilde{\omega}^t, \widetilde{\omega}^a, \widetilde{\omega}^v\}$. The formula is defined as follows:

$$\widetilde{\omega}^\gamma = \frac{\omega^\gamma}{\omega^t + \omega^a + \omega^v}. \tag{7}$$

Finally, we obtain the feature representation ξ after the fusion of the three modalities according to the attention weight. The formula is defined as follows:

$$\xi = \xi_t * \widetilde{\omega}^t + \xi_a * \widetilde{\omega}^a + \xi_v * \widetilde{\omega}^v. \tag{8}$$

2) Weighted Multirelational Affective Interaction Graph: As shown in Fig. 3, we build a multirelational affective interaction graph that includes the relationships between speakers and heterogeneous elements extracted from events. In particular, we construct a dynamic graph structure that changes over time. To capture the heterogeneous information contained in different relations, we construct a weighted multirelational affective interaction graph $G = \{V, \aleph, W, \{\Re_r^{\omega}\}_{r=1}^R\}$ to associate the relationship between nodes. The node set V in the multirelational emotional interaction graph is a series of fused multimodal feature vectors. The edge $e_{ij} \in \aleph$ is composed of speaker relation or event relation between v_i and v_j. $\omega_{ij} \in W$ is the weight of the edge e_{ij}. $r \in \Re$ is an edge relation.

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 03,2025 at 01:07:51 UTC from IEEE Xplore. Restrictions apply.

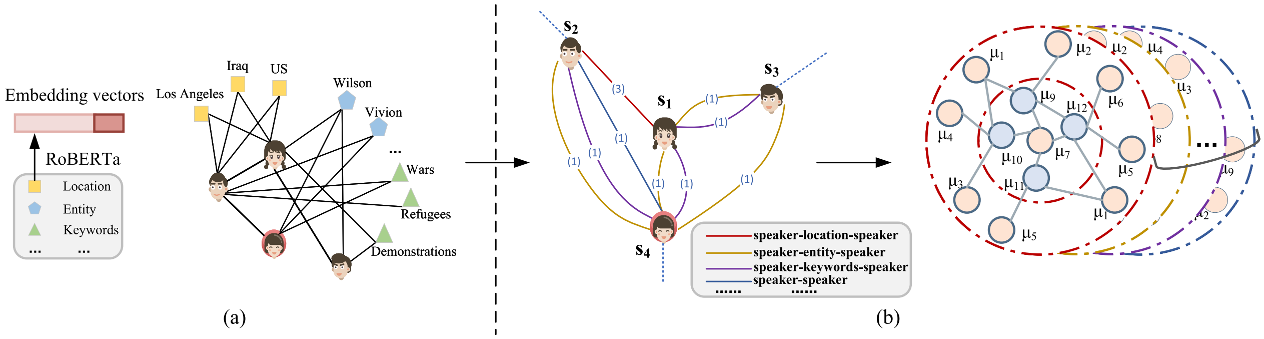


Fig. 3. (a) Heterogeneous dialog graph composed of dialog relations and event relations. (b) We split the heterogeneous graph to construct a weighted multirelational dialog graph.

The formula for the edge $e_{ij}^{r_E}$ of different relations composed of events is defined as follows:

$$e_{ij}^{r_E} = \min\left\{\left[A_{r_E} \cdot A_{r_E}^T\right]_{ij}, 1\right\} \tag{9}$$

where A_{r_E} represents the adjacency matrix of the multirelationship graph, where its rows represent all event nodes, and its columns represent event nodes belonging to relation r_E. $A_{r_E}^T$ is the transposition of the matrix A_{r_E}. To capture the difference between different edges under the same relationship, we define the weight of the edge $e_{ij}^{r_E}$ as follows:

$$\omega_{ij}^{r_E} = \left[A_{r_E} \cdot A_{r_E}^T\right]_{ij}. \tag{10}$$

For the edge e_{ij}^{rs} composed of the relationship between the speakers, if there is a dialog between the speakers, we connect an edge for them. Otherwise, no edge is established. For the edge weight ω_{ij}^{rs} of edge e_{ij}^{rs}, we use the similarity attention mechanism to assign weights to it. First, we use two linear layers to compute the similarity between nodes in the graph. The formula is defined as follows:

$$\rho_{ij}^{rs} = W_{\varpi_1}^{rs}\left(\text{ReLU}\left(W_{\varpi_2}^{rs}\left[\xi_i^{rs} \oplus \xi_j^{rs} \otimes \Im_{ij}\right] + b_2\right) + b_1\right) \tag{11}$$

where $W_{\varpi_1}^{rs}$ and $W_{\varpi_2}^{rs}$ are the learnable parameters in the linear layer. b_1 and b_2 are the biases of the linear layers. $\oplus$ means splicing, and $\otimes$ means dot multiplication. $\Im_{ij} \in \{0, 1\}$ and $\Im_{ij} = 1$ indicate that there is an edge connection between node i and node j, and $\Im_{ij} = 0$ indicates that there is not an edge connection between node i and node j.

Then, we use the attention mechanism to get the weight of each edge, and the formula is defined as follows:

$$\omega_{ij}^{rs} = \text{softmax}\left(\rho_{ij}^{rs}\right) = \frac{\exp\left(\rho_{ij}^{rs}\right)}{\sum_{n \in \mathcal{M}_i} \exp\left(\rho_{im}^{rs}\right)} \tag{12}$$

where $\mathcal{M}_i$ is the set of neighbor nodes of node i. The larger the ω_{ij}, the higher the correlation between nodes.

3) Self-Supervised Masked Graph Autoencoder: To improve the joint representation ability of features and structures of GNNs, we propose an SMGAE, which learns better feature embedding representation by randomly masking and reconstructing the nodes and edges in the graph. Unlike recent studies that only reconstruct features or structures, we reconstruct both features and structures to improve the generalization performance of the model.

First, we sample some nodes and edges in the graph and use the mask token to mask the node's feature vector and edge weights. Specifically, we use the Bernoulli distribution to generate a 0–1 matrix and then perform a dot product operation on the generated 0–1 matrix with the original feature matrix and adjacency matrix. Through the above operations, we can get the masked node set and edge set. The node feature formula after masking is defined as follows:

$$\tilde{\xi}_i = \begin{cases} \xi_{[M]}, & v_i \in V_M \\ \xi_i, & v_i \notin V_M \end{cases} \tag{13}$$

where V_M represents the masked node set and $\xi_{[M]}$ is the masked multimodal feature vector.

The formula for the masked edge is defined as follows:

$$\tilde{e}_{ij} = \begin{cases} e_{ij}^{[M]}, & \aleph_i \in \varphi_M \\ e_{ij}, & \aleph_i \notin \varphi_M \end{cases} \tag{14}$$

where φ_M represents the masked edge set and $e_{ij}^{[M]}$ represents the masked edge.

The goal of SMGAE is to reconstruct the masked node features and adjacency matrix A by using a small number of node features and edge weights. This article has adopted a graph convolutional neural network (GCN) as our encoder to aggregate information. The formula is defined as follows:

$$\begin{aligned} &p_\vartheta\left(\xi_i, e_i \mid \hat{\xi}_i, \hat{e}_i\right) \\ &= \sum_M p_\vartheta\left(\xi_i, e_i^{[M]} \mid e_i^{\sim[M]}, \hat{\xi}_i, \hat{e}_i\right) \cdot p_\vartheta\left(e_i^{[M]} \mid \hat{\xi}_i, \hat{e}_i\right) \\ &= \mathbb{E}_{[M]}\left[p_\vartheta\left(\xi_i, e_i^{\sim[M]} \mid e_i^{[M]}, \hat{\xi}_i, \hat{e}_i\right)\right] \\ &= \mathbb{E}_{[M]}\left[p_\vartheta\left(\xi_i \mid e_i^{[M]}, \hat{\xi}_i, \hat{e}_i\right) \cdot p_\vartheta\left(e_i^{\sim[M]} \mid e_i^{[M]}, \hat{\xi}_i, \hat{e}_i\right)\right] \end{aligned} \tag{15}$$

where $p_\vartheta(\xi_i \mid e_i^{[M]}, \hat{\xi}_i, \hat{e}_i)$ is the expected value of the generated node feature and $p_\vartheta(e_i^{\sim[M]} \mid e_i^{[M]}, \hat{\xi}_i, \hat{e}_i)$ is the expected value of the generated edge. $e_i^{[M]}$ is the unmasked edge. $\hat{\xi}_i$ and $\hat{e}_i$ represent the node features and edges generated by encoding, respectively.

In this article, we will use a GCN as our encoder to aggregate information, and the formula is defined as follows:

$$\begin{aligned} I_i^{(t)} = \text{ReLU}\Bigg(&\sum_{k \in \aleph_i^r} \sum_{r \in \Re} \sum_{j \in \mathcal{M}_i^r} \Bigg(\frac{w_{ij}^r}{c_{i,r}} W_r^{(t)} I_j^{(t-1)} \\ &+ w_{ii}^r W_\zeta^{(t)} I_i^{(t-1)}\Bigg) \cdot \tilde{e}_{ik}\Bigg) \end{aligned} \tag{16}$$

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 03,2025 at 01:07:51 UTC from IEEE Xplore. Restrictions apply.

where $I_i^{(t)}$ is the feature vector representation of node i at time t. $\aleph_i^r$ represents the edge set of node i under the edge relation $r \in \{\Re\}_{r=1}^{R}$. $\tilde{e}_{ik} \in [0, 1]$, $c_{i,r} = \|\mathcal{M}_i^r\|$. $W_\zeta^{(t)}$ is a learnable network parameter.

After getting the encoded feature vector, we need to use the decoder to map the latent feature distribution to the input ξ. The design of the encoder determines the ability of feature recovery, while simple decoders (such as MLPs) are less capable and cannot recover high-level semantic information. In this article, we choose the graph attention network (GAT) with stronger decoding ability as the decoder of SMGAE, which can utilize the surrounding neighbor information to recover the input features instead of just relying on the nodes themselves.

In the process of coding and decoding, we do not use the mean square error (MSE). Because it is easily affected by the vector dimension and norm, it instead uses the cosine similarity error, which is more stable in the training process and guides the optimization direction of the model gradient. The formula is defined as follows:

$$\begin{aligned}\cos(\xi_i, Z_i) &= \frac{\sum_{m=1}^{N}(\xi_i \cdot Z_i)}{\sqrt{\sum_{m=1}^{N}(\xi_i)^2} \cdot \sqrt{\sum_{m=1}^{N}(Z_i)^2}} + \lambda\|W\|_F^2 \\ &= \sum_{m=1}^{N} \frac{\xi_i \cdot \left(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}\xi_i W\right)}{\sqrt{(\xi_i)^2 \cdot \left(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}\xi_i W\right)^2}} + \lambda\|W\|_F^2\end{aligned} \tag{17}$$

where $Z_i = \widetilde{D}^{-(1/2)}\tilde{A}\widetilde{D}^{-(1/2)}\xi_i W$ is the feature vector decoded by the GNN. $\widetilde{D}$ is the degree matrix of the node, and $\tilde{A}$ is the adjacency matrix of the node. λ is a hyperparameter, and $\|W\|_F^2$ is the weight decay coefficient of the model, which is used to improve the robustness of the model.

In this article, we define $\hat{A} = \widetilde{D}^{-(1/2)}\tilde{A}\widetilde{D}^{-(1/2)}$, and the loss function becomes

$$\begin{aligned}\mathcal{L}(W)^{\text{node}} &= \operatorname{tr}\left(\frac{\xi \cdot (\hat{A}\xi W)}{\sqrt{\xi^T\xi} \cdot \sqrt{\left[(\hat{A}\xi W)^T(\hat{A}\xi W)\right]}}\right) + \lambda\|W\|_F^2 \\ &= \operatorname{tr}\left(\frac{\xi\hat{A}\xi W}{\sqrt{\xi^T\xi}\sqrt{W^T\xi^T\hat{A}^T\hat{A}\xi W}}\right) + \lambda\|W\|_F^2\end{aligned} \tag{18}$$

where $A\text{tr}(\cdot)$ is the trace of the matrix. Then, we can get the first-order partial derivative of $\mathcal{L}$ to W and set the value of the first-order partial derivative to 0 to obtain the optimal network parameter W. The formula is defined as follows:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial W} &= \frac{\xi\hat{A}\sqrt{\xi\xi^T}\sqrt{\xi W^T\xi^T\hat{A}^T\hat{A}\xi W}}{\xi\xi^T\xi W^T\xi^T\hat{A}^T\hat{A}\xi W} \\ &\quad - \frac{\xi\hat{A}\xi W\sqrt{\xi^T\xi}\,\frac{\sqrt{W^T\xi^T\hat{A}^T\hat{A}\xi}}{\sqrt{W}}}{\xi\xi^T\xi W^T\xi^T\hat{A}^T\hat{A}\xi W} + 2\lambda W \\ &= 0.\end{aligned} \tag{19}$$

For the reconstruction of the edge structure, we will use the contrastive loss of positive and negative samples to optimize,

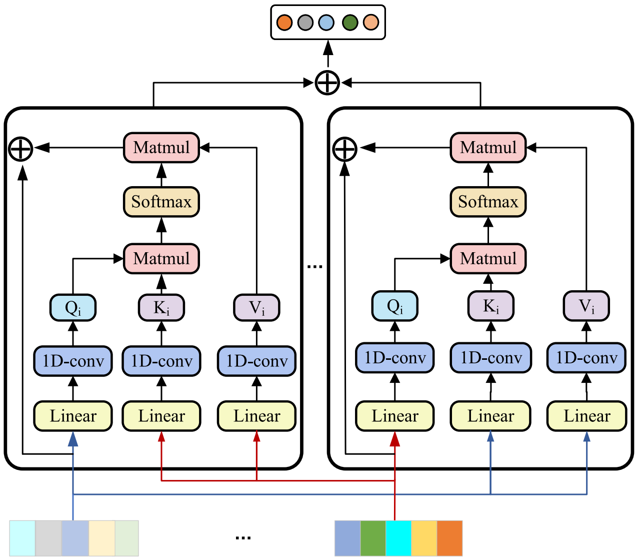


Fig. 4. MIT consists of multiple Transformer modules, each containing multiple linear layers, 1D-Conv, and softmax layers. MIT captures the underlying joint distribution between different relations by transferring information.

and the formula is defined as follows:

$$\mathcal{L}_i^{\text{edge}} = -\sum_{\varkappa^+ \in e_i^{[M]}} \log \frac{\exp\left(D_i^{\text{edge}}, D_{\varkappa^+}^{\text{edge}}\right)}{\sum_{j \in M_i^- \cup \{\varkappa^+\}} \exp\left(D_i^{\text{edge}}, D_{\varkappa}^{\text{edge}}\right)} \tag{20}$$

where $\varkappa^+$ represents the masked edge and D_i^{edge} represents the probability of the edge belonging to the ith node.

4) Weighted Relation-Aware Multiple Subgraph Information Aggregation: To better fuse the multiple information between relations and capture the correlation between different relations, we design an MIT to aggregate the interactive information between different relations through multiple information fusion. After modeling the information aggregation of multiple subgraphs, the sentiment classification effect of DER-GCN will be more credible.

As shown in Fig. 4, MIT is composed of Transformers with multiple cross branches, and the interactive information between different relations will be bidirectionally transmitted in MIT. Specifically, we first input the feature vectors obtained after MGAE learning into three fully connected layers and 1-D convolutional layers, respectively, to obtain vectors Q, K, and V. The formula is defined as follows:

$$\begin{aligned}\left[Q_i^1, Q_i^2, \ldots, Q_i^N\right] &= \operatorname{Conv}\left(\left[I_i^1, I_i^2, \ldots, I_i^N\right] W_Q^{\mathbb{R}^{d_I}}\right) \\ \left[K_i^1, K_i^2, \ldots, K_i^N\right] &= \operatorname{Conv}\left(\left[I_i^1, I_i^2, \ldots, I_i^N\right] W_K^{\mathbb{R}^{d_I}}\right) \\ \left[V_i^1, V_i^2, \ldots, V_i^N\right] &= \operatorname{Conv}\left(\left[I_i^1, I_i^2, \ldots, I_i^N\right] W_V^{\mathbb{R}^{d_I}}\right)\end{aligned} \tag{21}$$

where $W_Q^{\mathbb{R}^{d_I}}$, $W_K^{\mathbb{R}^{d_I}}$, and $W_V^{\mathbb{R}^{d_I}}$ are the learnable network parameters in the fully connected layer, and Conv is a 1-D convolution operation. Next, we use the softmax function to obtain the attention scores for feature vectors composed of different relations as follows:

$$\begin{aligned}&\left[\text{att}_{\text{score}}^1, \text{att}_{\text{score}}^2, \ldots, \text{att}_{\text{score}}^N\right]_i \\ &\quad = \operatorname{softmax}\left(\frac{\left[Q_i^1, Q_i^2, \ldots, Q_i^N\right]\left[K_i^1, K_i^2, \ldots, K_i^N\right]^T}{\varepsilon}\right)\end{aligned} \tag{22}$$

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 03,2025 at 01:07:51 UTC from IEEE Xplore. Restrictions apply.

where ε is the dimension of the feature vector Q. T represents the transposition of the matrix. Finally, we perform information fusion across relations by the following formula:

$$\hat{I}_i^{\vartheta} = I_i^{\vartheta} + \left[\text{att}_{\text{score}}^{1}, \text{att}_{\text{score}}^{\vartheta-1}, \ldots, \text{att}_{\text{score}}^{\vartheta+1}, \text{att}_{\text{score}}^{N}\right]_i \times \left[V_i^1, V_i^{\vartheta-1}, \ldots, V_i^{\vartheta+1}, V_i^N\right] \quad (23)$$

where ϑ represents the ϑth relation. After cross-relational information fusion, we can obtain multirelational fusion vectors containing rich semantic information.

5) Balanced Sampling Strategy-Based Contrastive Learning Mechanism: The number of emotions in each category in the multimodal emotion recognition in conversation (MERC) task is quite different. If the cross-entropy loss function is used to guide the learning process of the model, it will cause the model to have a serious overfitting effect on the minority category of emotions. Inspired by contrastive learning, it can learn discriminative boundary information for instances between classes. Therefore, it effectively alleviates the long-tailed problem in MERC.

Based on the above research, we introduce a triplet loss function in the process of model training to solve the problem of class distribution imbalance. In addition, we also add a global cross-entropy loss to preserve as much graph structure information as possible.

For each utterance m_i, we sample its positive samples m_i^+ and negative samples m_i^- to get the triplet loss value of the model, which narrows the gap between positive samples and actual samples. It can widen the gap between negative samples and actual samples. The formula is defined as follows:

$$\mathcal{L}_E = \sum_{\left(\varkappa_{m_i}, \varkappa_{m_i}^{+}, \varkappa_{m_i}^{-}\right)\in S} \max\left\{E\left(\varkappa_{m_i}, \varkappa_{m_i}^{+}\right) - E\left(\varkappa_{m_i}, \varkappa_{m_i}^{-}\right) + b, 0\right\} \quad (24)$$

where $E(,)$ is used to calculate the Euclidean distance between two feature vectors. b is a hyperparameter of the model that measures the distance between samples.

We also construct a global cross-entropy loss to preserve the information of similar structures better. The formula is defined as follows:

$$\mathcal{L}_C = -\frac{1}{\sum_{m=1}^{\sigma} \mathcal{L}_i} \sum_{m=1}^{\sigma} \sum_{n=1}^{\gamma_n} \sum_{k=1}^{\lambda} y_{m,k}^{n} \log_2\left(\hat{y}_{m,k}^{n}\right) \quad (25)$$

where θ is the total number of dialogs in the benchmark dataset, γ_n represents the number of utterances in the nth dialog, and λ is the total number of sentiment categories.

6) Emotion Classification: The emotional features E_f obtained after going through the GCN are sent to a linear layer with residual connections and then go through a layer of softmax layer to obtain the probability distribution P of emotional labels: the formula is defined as follows:

$$\begin{aligned} \alpha &= E_f + \text{ReLU}\left(E_f W_f + b_f\right) \\ P &= \text{softmax}\left(\alpha W_\alpha + b_\alpha\right) \end{aligned} \quad (26)$$

where $W_f \in R^{d_f \times d_f}$, $b_f \in \mathbb{R}^{d_f}$, $W_\alpha \in \mathbb{R}^{d_f \times d_\lambda}$, and $b_\alpha \in \mathbb{R}^{\lambda}$ are parameters that can be learned in the model.

We get the sentiment label with the maximum probability through the argmax function

$$\widehat{y} = \text{argmax}\ (P) \quad (27)$$

where $\hat{y}$ represents the sentiment label predicted by the model.

V. Experiments

A. Benchmark Dataset Used

The IEMOCAP [28] and MELD [29] benchmark datasets are two multimodal dialog sentiment datasets that researchers widely use to evaluate the effectiveness of their models.

The IEMOCAP database is a multimodal emotion recognition dataset. The IEMOCAP dataset contains three modalities of the speaker's video, audio, and dialog text. The dataset contains five actors and five actresses, and each dialog scene has a dialog between an actor and an actress. The labels of these conversations are all manually annotated, and at least three experts in the emotion domain are assigned to each conversation.

The MELD is a popular multimodal benchmark dataset in the MDER domain, consisting of multiple dialog clips from the TV series Friends. The total video and audio duration of MELD is approximately 13.7 h, and each video clip contains multiple speakers. The labels of these conversations are all manually annotated, and at least five experts in the emotion domain are assigned to each conversation.

B. Evaluation Metrics

In this section, we illustrate the evaluation metrics used to verify the effectiveness of the model proposed in this article. This article uses the following four evaluation metrics: 1) accuracy; 2) $F1$; 3) WA; and 4) weight $F1$ (WF1). Due to the serious data imbalance problem in the IEMOCAP and MELD benchmark datasets, we will mainly use WA and WF1 as our main evaluation metrics.

C. Baseline Models

To verify the effectiveness of our model on the IEMOCAP and MELD benchmark datasets, we conduct comparative experiments with 12 state-of-the-art deep learning (DL)-based algorithms, including one traditional CNN algorithm (i.e., TextCNN [30]), four RNN algorithms (i.e., bidirectional contextual LSTM (bc-LSTM) [31], DialogueRNN [32], conversational memory network (CMN) [33], and adapted dynamic memory network (A-DMN) [34]), three GNN algorithms (i.e., DialogueGCN [35], relation-aware graph attention networks (RGAT) [36], and LR-GCN [37]), one feature fusion algorithm (i.e., low-rank multimodal fusion (LFM) [38]), and three pretrained algorithms (context modeling with speaker's pre-trained memory (CoPMP) [39], EmoBERTa [40], and COGMEN [41]).

VI. Results and Discussion

A. Comparison With Baselines

To verify the effectiveness of the DER-GCN model proposed in this article, we have done extensive experiments to compare it with other comparison algorithms. Tables I and II present the emotion recognition effects of DER-GCN and other comparative algorithms on two popular datasets, respectively.

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 03,2025 at 01:07:51 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPARISON WITH OTHER BASELINE MODELS ON THE IEMOCAP DATASET. ACC. = ACCURACY. AVERAGE(w) = WEIGHTED AVERAGE

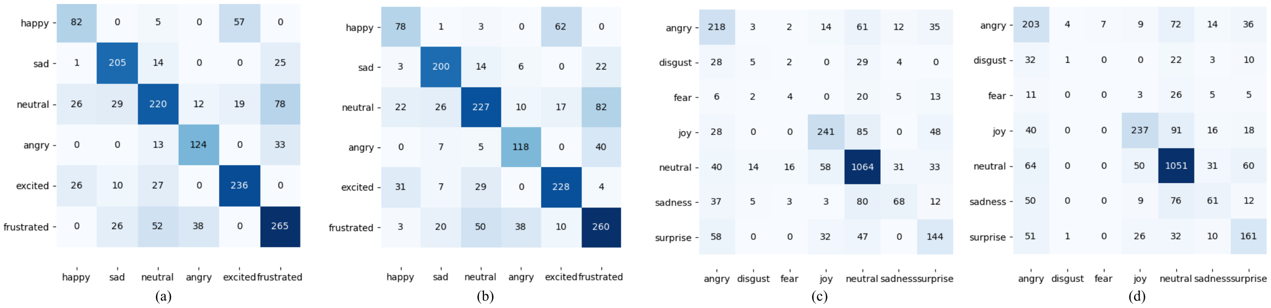
Methods	IEMOCAP													
	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average(w)	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TextCNN	27.7	29..8	57.1	53.8	34.3	40.1	61.1	52.4	46.1	50.0	62.9	55.7	48.9	48.1
bc-LSTM	29.1	34.4	57.1	60.8	54.1	51.8	57.0	56.7	51.1	57.9	67.1	58.9	55.2	54.9
CMN	25.0	30.3	55.9	62.4	52.8	52.3	61.7	59.8	55.5	60.2	**71.1**	60.6	56.5	56.1
LFM	25.6	33.1	75.1	78.8	58.5	59.2	64.7	65.2	80.2	71.8	61.1	58.9	63.4	62.7
A-DMN	43.1	50.6	69.4	76.8	63.0	62.9	63.5	56.5	**88.3**	**77.9**	53.3	55.7	64.6	64.3
DialogueGCN	40.6	42.7	**89.1**	**84.5**	62.0	63.5	67.5	64.1	65.5	63.1	64.1	66.9	65.2	64.1
RGAT	60.1	51.6	78.8	77.3	60.1	65.4	70.7	63.0	78.0	68.0	64.3	61.2	65.0	65.2
CoMPM	59.9	60.7	78.0	82.2	60.4	63.0	70.2	59.9	85.8	78.2	62.9	59.5	67.7	67.2
EmoBERTa	56.9	56.4	79.1	83.0	64.0	61.5	70.6	69.6	86.0	78.0	63.8	68.7	67.3	67.3
COGMEN	57.4	51.9	81.4	81.7	65.4	**68.6**	69.5	66.0	83.3	75.3	63.8	68.2	68.2	67.6
LR-GCN	54.2	55.5	81.6	79.1	59.1	63.8	69.4	69.0	76.3	74.0	68.2	**68.9**	68.5	68.3
DER-GCN	**60.7**	**58.8**	75.9	79.8	**66.5**	61.5	**71.3**	**72.1**	71.1	73.3	66.1	67.8	**69.7**	**69.4**

TABLE II
COMPARISON WITH OTHER BASELINE MODELS ON THE MELD DATASET. ACC. = ACCURACY. AVERAGE(w) = 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	Acc.	F1
TextCNN	76.2	74.9	43.3	45.5	4.6	3.7	18.2	21.1	46.1	49.4	8.9	8.3	35.3	34.5	56.3	55.0
bc-LSTM	78.4	73.8	46.8	47.7	3.8	5.4	22.4	25.1	51.6	51.3	4.3	5.2	36.7	38.4	57.5	55.9
DialogueRNN	72.1	73.5	54.4	49.4	1.6	1.2	23.9	23.8	52.0	50.7	1.5	1.7	41.0	41.5	56.1	55.9
DialogueGCN	70.3	72.1	42.4	41.7	3.0	2.8	20.9	21.8	44.7	44.2	6.5	6.7	39.0	36.5	54.9	54.7
RGAT	76.0	78.1	40.1	41.5	3.0	2.4	32.1	30.7	68.1	58.6	4.5	2.2	40.0	44.6	60.3	61.1
CoMPM	78.3	82.0	48.3	49.2	1.7	2.9	35.9	32.3	71.4	61.5	3.1	2.8	42.2	45.8	64.1	65.3
EmoBERTa	**78.9**	**82.5**	50.2	50.2	1.8	1.9	33.3	31.2	**72.1**	61.7	9.1	2.5	43.3	46.4	64.1	65.2
A-DMN	76.5	78.9	**56.2**	**55.3**	8.2	8.6	22.1	24.9	59.8	57.4	1.2	3.4	41.3	40.9	61.5	60.4
LR-GCN	76.7	80.0	53.3	55.2	0.0	0.0	49.6	35.1	68.0	**64.4**	10.7	2.7	48.0	51.0	65.7	65.6
DER-GCN	76.8	80.6	50.5	51.0	**14.8**	**10.4**	**56.7**	**41.5**	69.3	64.3	**17.2**	**10.3**	**52.5**	**57.4**	**66.8**	**66.1**

1) IEMOCAP: As shown in Table I, compared with other comparison algorithms, our proposed MDER method DER-GCN has the best emotion recognition effect on the IEMOCAP dataset, and the WA and WF1 values are 69.7% and 69.4%, respectively. DER-GCN proposes a method for dialog emotion recognition that comprehensively considers sequential context information, dialog relations between speakers, and event relations. Among other benchmark models, latent relation-aware graph convolutional network (LR-GCN) performs slightly worse than DER-GCN, with the WA and WF1 values of 68.5% and 68.3%, respectively. We speculate that LR-GCN outperforms other baseline models, because it considers both the interaction between speakers and the latent semantic relationship of the dialog context. However, LR-GCN ignores the event relations in the dialog, so its emotion recognition effect is lower than that of the model proposed in this article, DER-GCN. The emotion recognition effect of CoMPM, EmoBERTa, and CoMPM is lower than DER-GCN and LR-GCN. Similar to LR-GCN, they both ignore the impact of event relations in the dialog on emotion recognition. The emotion recognition effect of DialogueGCN and RGAT is only about 65%, which is because they only consider the influence of the dependency relationship between speakers or the position information of sequential context on emotion recognition. The emotion prediction effect of A-DMN and LFM is much lower than that of DER-GCN, with the WA values of 64.6% and 63.4% and the WF1 values of 64.3% and 62.7%, respectively. It is because they do not model speaker relations and event relations in dialog, although they design a fusion mechanism to obtain complementary multimodal semantic features. The emotion prediction performance of other baseline methods, such as TextCNN, is much worse than that of DER-GCN, because they only model sequential context information, which results in limited semantic information learned by the model. Overall, DER-GCN outperforms other baselines in accuracy on "happy," "neutral," and "angry," and DER-GCN outperforms other baselines in $F1$ on "happy" and "angry." In addition, DER-GCN is also less far behind the baseline on other emotion categories.

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 03,2025 at 01:07:51 UTC from IEEE Xplore. Restrictions apply.



	happy	sad	neutral	angry	excited	frustrated
happy	82	0	5	0	57	0
sad	1	205	14	0	0	25
neutral	26	29	220	12	19	78
angry	0	0	13	124	0	33
excited	26	10	27	0	236	0
frustrated	0	26	52	38	0	265

(a)

	happy	sad	neutral	angry	excited	frustrated
happy	78	1	3	0	62	0
sad	3	200	14	6	0	22
neutral	22	26	227	10	17	82
angry	0	7	5	118	0	40
excited	31	7	29	0	228	4
frustrated	3	20	50	38	10	260

(b)

	angry	disgust	fear	joy	neutral	sadness	surprise
angry	218	3	2	14	61	12	35
disgust	28	5	2	0	29	4	0
fear	6	2	4	0	20	5	13
joy	28	0	0	241	85	0	48
neutral	40	14	16	58	1064	31	33
sadness	37	5	3	3	80	68	12
surprise	58	0	0	32	47	0	144

(c)

	angry	disgust	fear	joy	neutral	sadness	surprise
angry	203	4	7	9	72	14	36
disgust	32	1	0	0	22	3	10
fear	11	0	0	3	26	5	5
joy	40	0	0	237	91	16	18
neutral	64	0	0	50	1051	31	60
sadness	50	0	0	9	76	61	12
surprise	51	1	0	26	32	10	161

(d)

Fig. 5. Classification of DER-GCN and LR-GCN on the IEMOCAP and MELD datasets. (a) Confusion matrix obtained by DER-GCN on the IEMOCAP dataset. (b) Confusion matrix obtained by LR-GCN on the IEMOCAP dataset. (c) Confusion matrix obtained by DER-GCN on the MELD dataset. (d) Confusion matrix obtained by LR-GCN on the MELD dataset.

2) MELD: As shown in Table II, the emotion prediction effect of the DER-GCN model on the MELD dataset is better than other comparison algorithms, and the WA and WF1 values are 66.8% and 66.1%, respectively. The effect of LR-GCN is second, with the WA and WF1 values of 65.7% and 65.6%, respectively. The prediction performance of A-DMN is lower than that of DER-GCN and LR-GCN, with the WA and WF1 values of 61.5% and 60.4%, respectively. Other comparison algorithms perform poorly, because they all ignore modeling the relationships between speakers. In addition, compared with other comparison algorithms, DER-GCN has significantly improved the prediction accuracy on the minority class sentiment labels "fear" and "disgust." Specifically, the WA and WF1 values of DER-GCN on the "fear" label are 14.8% and 10.4%, respectively, and the prediction effect is improved by about 10%. The WA and WF1 values of DER-GCN on the "disgust" label are 17.2% and 10.3%, respectively, and the prediction effect is improved by about 10%. We have guessed that DER-GCN can improve the prediction effect of minority class sentiment. The model adopts a loss optimization strategy based on the contrastive learning mechanism, which can better represent minority class features. Overall, DER-GCN outperforms other baselines in accuracy and $F1$ on "fear," "sadness," "anger," and "disgust." In addition, DER-GCN is also less far behind the baseline on other emotion categories.

The experimental results show that the event relationship in the dialog significantly strengthens the model's understanding of the speaker's emotion. In addition, cross-modal feature fusion and loss optimization strategy based on contrastive learning can also enhance the model's emotion classification ability.

B. Analysis of the Experimental Results

To clarify the feature representation ability of the model on each emotion category, we analyze the distribution of emotion classification of DER-GCN and LR-GCN on the test set. Fig. 5 presents the confusion matrix for emotion classification by DER-GCN and LR-GCN on the IEMOCAP and MELD datasets.

Overall, our model DER-GCN predicts more correctly than LR-GCN on different emotion categories on the IEMOCAP and MELD datasets. For example, DER-GCN correctly predicts 1064 data on neutral emotion, while LR-GCN correctly predicts 1051 data on the MELD dataset. Therefore, the predictive ability of DER-GCN is better than LR-GCN. The performance improvement may be attributed to the ability of DER-GCN to learn the corresponding sentiment categories from events.

On the IEMOCAP dataset, we observe the confusion matrix and find that the DER-GCN easily misclassifies a "neutral" sentiment into a "frustrated" and "sad" sentiment. We believe that this is because there are semantically similar parts between "neutral" sentiment and "frustrated" or "sad" sentiment, which leads to fuzzy class boundaries in the representation of emotional features among different categories learned by DER-GCN. At the same time, we also find that the model incorrectly classified a "frustrated" or "sad" sentiment as a "neutral" sentiment. In addition, DER-GCN also has a mutual misclassification between "sad" sentiment and "frustrated" sentiment. There is also overlapping semantic information between "happy" sentiment and "excited" sentiment, which makes it difficult for DER-GCN to distinguish these emotions. The classification effect of the model in "sad" or "angry" sentiment is relatively good. Most of the tested utterances can be correctly classified. For the "excited" sentiment, we find that the DER-GCN misclassifies it as the "sad" sentiment. We think this is because speakers usually express their emotions more implicitly and sarcastically when they talk about sensitive topics, and the model cannot capture this semantic information.

The MELD dataset shows a specific semantic correlation between the "neutral" sentiment and other types of emotion. Therefore, DER-GCN is prone to misclassify the "neutral" sentiment as other emotions. The opposite is also true. For the "surprise" sentiment, DER-GCN incorrectly classifies it into "joy" and "anger" sentiment. We guess this is because speakers with "surprise" sentiments are usually accompanied by "joy" or "anger" sentiments.

On the one hand, the speaker is stimulated by something wrong to produce surprise-like emotions, which will cause the speaker to feel angry. On the other hand, the speaker is surprised by the surprise prepared by others, which will cause the speaker to feel joy. For the "fear" sentiment, the model is prone to misclassify it as the "neutral" sentiment. For the "disgust" sentiment, the number of test utterances correctly classified by DER-GCN is minimal, and the classification results are unreliable. This is because the number of "disgust" sentiments in the MELD dataset is very small. DER-GCN cannot learn effective semantic information from such a small amount of data. At the same time, this problem also exists in the "fear" category of emotions. For the "angry" sentiment,

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 03,2025 at 01:07:51 UTC from IEEE Xplore. Restrictions apply.

TABLE III

Effect of DER-GCN on Two Datasets Using Unimodal Features and Multimodal Features, Respectively. T, V, and A Represent Text, Video, and Audio Modality Features

Modality	IEMOCAP		MELD	
	WA	WF1	WA	WF1
T	63.2	63.8	62.8	61.9
A	61.4	61.6	62.1	61.3
V	57.8	57.1	60.5	60.6
T+A	65.8	64.7	63.8	62.6
T+V	64.4	64.0	63.1	63.4
V+A	61.2	60.9	60.3	59.8
T+A+V	**69.7**	**69.4**	**66.8**	**66.1**

DER-GCN not only misclassifies it as the "surprise" sentiment but also misclassifies it as the "sadness" or the "joy" sentiment. On the one hand, speakers with "angry" emotions are usually accompanied by "sadness" sentiments. On the other hand, speakers with an "angry" sentiment may be more implicit in expressing their emotions. The above two reasons may lead to biases in DER-GCN in understanding the semantics of test utterances.

C. Importance of the Modalities

To verify the importance of the three modal features of text, video, and audio, we conduct experiments on the IEMOCAP and MELD datasets to compare the performance of unimodal, bimodal, and multimodal features. The experimental results are shown in Table III. Due to the problem of data imbalance in the dataset, WF1 comprehensively considers the precision rate and recall rate. So, we chose WF1 as our main evaluation metric and WA as our secondary evaluation metric. For the experimental results of the single modality, the values of WA and WF1 of the text modality are higher than the audio and video modality. The values of WA are 63.2% and 62.8% on the IEMOCAP and MELD datasets, respectively, and the values of WF1 are 63.8% and 61.9% on the IEMOCAP and MELD datasets, respectively, which indicates that the text modal features play the most important role in the emotion recognition of the model. The effect of the audio modality is second, the values of WA are 61.4% and 62.1%, respectively, and the values of WF1 are 61.6% and 61.3%, respectively. The video modality performs the worst, with the values of 57.8% and 60.5% for WA, respectively, and with the values of 57.1% and 60.6% for WF1, respectively, indicating that it is difficult for the model to extract useful emotional features from video features. The experimental results show that the noise introduced by the text features is the least, which will benefit the model in learning the embedded representation of the emotional features.

The experimental results of bimodality are better than single modality. The WA value is improved by 0.2%–8%, and the WF1 value is improved by 0.7%–7%. It indicates that emotional features are not only related to contextual information but also changes in sound signals in audio and facial expressions in video. The bimodal feature combines two different unimodal features, which can effectively improve the emotion prediction effect of the model. Furthermore, the bimodal features fused with text and audio performed the best emotion prediction, with the values of 65.8% and 63.8% for WA, respectively, and with the values of 64.7% and 62.6% for WF1, respectively. The emotion prediction effect of bimodal features fused by text and video is second, with the values of 64.4% and 63.1% for WA, respectively, and with the values of 64.0% and 63.4% for WF1, respectively. The bimodal features fused with audio and video have the worst emotion prediction performance, with the values of 61.2% and 60.3% for WA, respectively, and with the values of 60.9% and 59.8% for WF1, respectively.

After the fusion of three modal features of text, video, and audio, the multimodal features have the best emotion prediction performance. It is better than the performance of single-modal and bimodal features, which indicates that the model not only utilizes the semantic information of the dialog context but also utilizes video and audio features to enhance the representation ability of the emotional feature vectors.

TABLE IV

Emotion Recognition Effects of Different Multimodal Feature Fusion Methods on IEMOCAP and MELD Datasets. We Use Three Modal Text, Video, and Audio Features for Each Method

Methods	IEMOCAP		MELD	
	WA	WF1	WA	WF1
Add	65.2	64.8	62.9	62.4
Concatenate	64.6	64.1	62.5	61.6
Tensor Fusion	66.7	65.6	63.7	63.5
Cross-modal Fusion(Ours)	**69.7**	**69.4**	**66.8**	**66.1**

D. Effectiveness of Cross-Modal Feature Fusion

In this section, to verify the effectiveness of the cross-modal feature fusion method proposed in this article, we compare it with the other three fusion methods, i.e., add and concatenation operation and tensor fusion network (TFN).

The experimental results are shown in Table IV. Compared with other multimodal feature fusion methods, the cross-modal feature fusion method proposed in this article has achieved the best experimental results. The values of WA are 69.7% and 66.8%, respectively, and the values of WF1 are 69.4% and 66.1%, respectively. Specifically, compared with the add method, the WA value of the cross-modal feature fusion method is improved by 3.9%–4.5%, and the WF1 value is improved by 3.7%–4.6%. We think that the add method cannot capture the complementary semantic information between different modalities. The cross-modal feature fusion method can extract the most relevant semantic information with emotional features through the attention mechanism, thereby improving the emotion recognition effect of the model. At the same time, compared with the concatenate method, the WA value of the cross-modal feature fusion method is improved by 4.3%–5.1%, and the WF1 value is improved by 4.5%–5.3%. The reason is that the feature vector dimensions of the text, video, and audio modalities are high, leading to the combinational explosion of multimodal embedding representations generated by feature concatenation. Different from the

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 03,2025 at 01:07:51 UTC from IEEE Xplore. Restrictions apply.

TABLE V

DIFFERENT CONTEXT MODELING METHODS ON THE TWO DATASETS. ALL METHODS HAVE EXPERIMENTED WITH MULTIMODAL FEATURES

Methods	IEMOCAP		MELD	
	WA	WF1	WA	WF1
Without contextual modeling	62.3	61.7	60.1	61.6
Uni-GRU	67.1	66.2	63.4	63.0
Bi-GRU(Ours)	**69.7**	**69.4**	**66.8**	**66.1**

concatenate method, the cross-modal feature fusion method can achieve efficient feature dimensionality reduction while capturing rich semantic information. In addition, compared with the tensor fusion method, the WA value of the cross-modal feature fusion method is improved by 3%–3.1%, and the WF1 value is improved by 2.6%–3.8%. This is because the tensor fusion method needs to use tensors for feature representation, which introduces much computational consumption and reduces emotion recognition accuracy. The above experimental results demonstrate the effectiveness of the cross-modal feature fusion method proposed in this article.

E. Effectiveness of Bi-GRU

To verify the effectiveness of Bi-GRU for contextual semantic information extraction, we use three methods for comparative experiments. The experimental results are shown in Table V.

1) *Without Contextual Modeling:* This method does not use any contextual information modeling method for emotion recognition. Specifically, we replace the GRU layers with linear layers.
2) *Unidirectional GRU (Uni-GRU):* Instead of modeling context information, we use a Uni-GRU to extract contextual semantic information, which can memorize utterance information before the current moment.
3) *Bidirectional GRU (Bi-GRU):* Different from the above methods, we use Bi-GRU to model two opposite contextual utterances, which contain richer contextual information.

Among the three contrasting methods, we find that the emotion recognition method that does not model contextual semantic information works the worst, with the WA values of 62.3% and 60.1% on IEMOCAP and MELD datasets, respectively, and with the WF1 values of 61.7% and 61.6%, indicating the necessity of contextual semantic information modeling. The Uni-GRU method outperforms methods that do not model contextual semantic information, with the values of 67.1% and 63.4% for WA, respectively, and with the values of 66.2% and 63.0% for WF1, respectively. Bi-GRU performs the best for emotion recognition, with the WA values of 69.7% and 66.8%, respectively, and with the WF1 values of 69.4% and 66.1%, respectively. Compared with the other two methods, the WA value is increased by 2.6%–7.4%, and the WF1 value is increased by 3.1%–7.7%. Therefore, the experimental results show that the emotional information of the current moment is related to both historical discourse and future discourse.

F. Effectiveness of SMAGE and MIT

To explore the influence of SMAGE and MIT on the effect of emotion recognition, we conducted ablation experiments on the IEMOCAP and MELD datasets. As shown in Table VI, the emotion recognition performance of DER-GCN (w/o S/M) is much lower than that of DER-GCN (S) and DER-GCN (M). In addition, DER-GCN performs the best for emotion recognition. Experimental results demonstrate the effectiveness of SMAGE and MIT.

TABLE VI

EXPERIMENTAL RESULTS OF THE INFLUENCE OF SMGAE AND MIT ON THE EFFECT OF EMOTION RECOGNITION. DER-GCN (S) MEANS ONLY USING SMAGE WITHOUT USING MIT. DER-GCN (S) MEANS ONLY USING MIT WITHOUT USING SMAGE

Methods	IEMOCAP	MELD
	WF1	WF1
DER-GCN (w/o S/M)	62.7	60.8
DER-GCN (S)	67.8	65.2
DER-GCN (M)	67.0	64.4
DER-GCN	69.4	66.1

TABLE VII

EXPERIMENTAL RESULTS OF THE DER-GCN METHOD FOR MINORITY EMOTION RECOGNITION TASKS (I.E., HAPPY, FEAR, AND DISGUST) ON THE IEMOCAP AND MELD DATASETS. WAF1 IS CHOSEN AS THE EVALUATION CRITERION FOR THE EXPERIMENTS. DER-GCN (W/O B) INDICATES THAT THE BALANCED SAMPLING STRATEGY-BASED CONTRASTIVE LEARNING MECHANISM IS NOT INTRODUCED

Methods	IEMOCAP	MELD	
	Happy	Fear	Disgust
DER-GCN (w/o B)	55.6	1.9	2.7
DER-GCN	58.8	10.4	10.3

G. Effectiveness of Balanced Sampling Strategy

To verify whether the balanced sampling strategy-based contrastive learning mechanism for long-tailed problems can improve the emotion recognition effect of minority emotions, we conduct an ablation experiment of the balanced sampling strategy-based contrastive learning mechanism on minority emotions (i.e., happiness, fear, and disgust). As shown in Table VII, the emotion recognition effect of DER-GCN (w/o B) on minority emotions is particularly poor, especially on fear and disgust. Compared with DER-GCN (B), DER-GCN has greatly improved the emotion recognition performance of minority emotions. The performance improvement may be attributed to the introduction of the balanced sampling strategy-based contrastive learning mechanism, which can optimize the feature representation of minority emotion.

H. Effectiveness of Event Graph

To verify the effectiveness of event graphs, we conduct ablation experiments on the impact of event graphs on emotion recognition. As shown in Table VIII, the emotion recognition performance of DER-GCN with an event graph is significantly better than that of DER-GCN (w/o E). Experiments demonstrate the effectiveness of event graphs.

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 03,2025 at 01:07:51 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII

EXPERIMENTAL RESULTS OF THE PROPOSED EVENT GRAPH FOR MULTIMODAL EMOTION RECOGNITION TASKS ON THE IEMOCAP AND MELD DATASETS. DER-GCN (E) INDICATES THAT USING AN EVENT GRAPH. DER-GCN (W/O E) MEANS USING A SPEAKER RELATION GRAPH WITHOUT USING AN EVENT GRAPH

Methods	IEMOCAP		MELD	
	WAA	WAF1	WAA	WAF1
DER-GCN (w/o E)	66.4	65.1	60.2	59.8
DER-GCN (E)	69.7	69.4	66.8	66.1

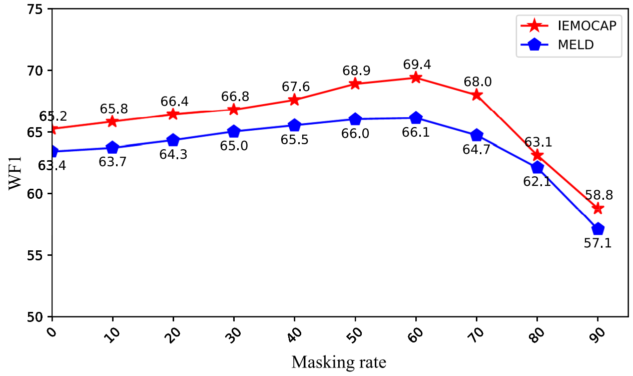


Fig. 6. Emotion recognition effect of DER-GCN under different node masking rates. DER-GCN achieves the best emotion recognition with a mask rate of about 60%.

I. Hyperparameter Settings

To verify the emotion recognition effect of DER-GCN under different mask rates, we conducted hyperparameter setting experiments on the IEMOCAP and MELD datasets. In particular, we find that DER-GCN works best for emotion recognition with an edge mask rate of 10%. However, the emotion recognition effect of the model is relatively poor under other edge mask rates. The intuition behind it is that if the edge mask rate is too large, the semantic information of the graph structure will be seriously lost, and the optimization ability of the graph structure is limited if the edge mask rate is too small. In addition, we also study the effect of different node masking rates on emotion recognition under the condition that the edge masking rate is 10%. As shown in Fig. 6, when the node masking rate is less than 60%, the emotional recognition effect of the model on the IEMOCAP and MELD datasets gradually increases. The model performs best in emotion recognition when the node mask rate is equal to 60%. When the node masking rate is less than 60%, the emotion recognition effect of the model begins to decline.

VII. CONCLUSION AND FUTURE WORK

This article proposes the DER-GCN model, which enables multimodal emotion recognition for multiple dialog relations. To capture the potential semantic information related to the dialog topic during the dialog process, we use an event extraction method to extract the main events in the dialog. In order to obtain better node embedding representation, we design a graph autoencoder based on node and edge masking mechanism, which reconstructs the original graph's topological structure and feature vectors through self-supervised learning. We introduce a sampling strategy based on contrastive learning to alleviate the data imbalance problem. DER-GCN is used to learn optimal network parameters in the multimodal emotion recognition task. On the IEMOCAP and MELD benchmark datasets, DER-GCN has greatly improved the effect of emotion recognition compared with other comparison algorithms.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and associate editor (AE) for their careful work and thoughtful suggestions that have helped improve this article substantially.

REFERENCES

[1] S. K. Khare and V. Bajaj, "Time–frequency representation and convolutional neural network-based emotion recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 2901–2909, Jul. 2021.

[2] L. Yi and M.-W. Mak, "Improving speech emotion recognition with adversarial data augmentation network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 172–184, Jan. 2022.

[3] Z. Lian, B. Liu, and J. Tao, "PIRNet: Personality-enhanced iterative refinement network for emotion recognition in conversation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 2863–2874, 2024.

[4] W. J. Baddar and Y. M. Ro, "Mode variational LSTM robust to unseen modes of variation: Application to facial expression recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 3215–3223.

[5] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 985–1000, 2021.

[6] Z. Zhang, P. Cui, J. Pei, X. Wang, and W. Zhu, "Eigen-GNN: A graph structure preserving plug-in for GNNs," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 2544–2555, Mar. 2023.

[7] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[8] T. Akilan, Q. J. Wu, A. Safaei, J. Huo, and Y. Yang, "A 3D CNN-LSTM-based image-to-image foreground segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 959–971, Mar. 2020.

[9] Y. Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31. Cambridge, MA, USA: MIT Press, 2018, pp. 4485–4495.

[10] S. Zheng, W. Cao, W. Xu, and J. Bian, "Revisiting the evaluation of end-to-end event extraction," in *Proc. Findings Assoc. Comput. Linguistics, ACL-IJCNLP*, 2021, pp. 4609–4617.

[11] Z. Hou et al., "GraphMAE: Self-supervised masked graph autoencoders," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2022, pp. 594–604.

[12] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1634–1654, 2023.

[13] W. Kong, M. Qiu, M. Li, X. Jin, and L. Zhu, "Causal graph convolutional neural network for emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 15, no. 4, pp. 1686–1693, 2023.

[14] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence LSTM architecture," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6474–6478.

[15] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition," *Inf. Process. Manage.*, vol. 57, no. 3, May 2020, Art. no. 102185.

[16] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jul. 2020, pp. 3507–3511.

[17] D. Sheng, D. Wang, Y. Shen, H. Zheng, and H. Liu, "Summarize before aggregate: A global-to-local heterogeneous graph inference network for conversational emotion recognition," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 4153–4163.

[18] Q. Huang et al., "Personalized dialogue generation with persona-adaptive attention," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 12916–12923.

[19] Y. Zheng, R. Zhang, M. Huang, and X. Mao, "A pre-training based personalized dialogue generation model with persona-sparse data," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 9693–9700.

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 03,2025 at 01:07:51 UTC from IEEE Xplore. Restrictions apply.

[20] Y. Zeng and J.-Y. Nie, "A simple and efficient multi-task learning approach for conditioned dialogue generation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 4927–4939.

[21] D. Liu, S. Xu, X.-Y. Liu, Z. Xu, W. Wei, and P. Zhou, "Spatiotemporal graph neural network based mask reconstruction for video object segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 3, pp. 2100–2108.

[22] S. Pan, R. Hu, G. Long, J. Jiang, L. Yao, and C. Zhang, "Adversarially regularized graph autoencoder for graph embedding," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 2609–2615.

[23] X. Wang, N. Liu, H. Han, and C. Shi, "Self-supervised heterogeneous graph neural network with co-contrastive learning," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 1726–1736.

[24] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21798–21809.

[25] D. Cai, S. Qian, Q. Fang, J. Hu, W. Ding, and C. Xu, "Heterogeneous graph contrastive learning network for personalized micro-video recommendation," *IEEE Trans. Multimedia*, vol. 25, no. 1, pp. 2761–2773, 2023.

[26] P. Peng, J. Lu, T. Xie, S. Tao, H. Wang, and H. Zhang, "Open-set fault diagnosis via supervised contrastive learning with negative out-of-distribution data augmentation," *IEEE Trans. Ind. Informat.*, vol. 19, no. 3, pp. 2463–2473, Mar. 2023.

[27] P. Gupta, "MERASTC: Micro-expression recognition using effective feature encodings and 2D convolutional neural network," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1431–1441, 2023.

[28] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[29] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 527–536.

[30] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empir. Method Nat. Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1746–1751.

[31] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 873–883.

[32] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 6818–6825.

[33] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, p. 2122.

[34] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1426–1439, Jul. 2022.

[35] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 154–164.

[36] T. Ishiwatari, Y. Yasuda, T. Miyazaki, and J. Goto, "Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 7360–7370.

[37] M. Ren, X. Huang, W. Li, D. Song, and W. Nie, "LR-GCN: Latent relation-aware graph convolutional network for conversational emotion recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 4422–4432, 2022.

[38] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[39] J. Lee and W. Lee, "CoMPM: Context modeling with speaker's pretrained memory tracking for emotion recognition in conversation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2022, pp. 5669–5679.

[40] T. Kim and P. Vossen, "EmoBERTa: Speaker-aware emotion recognition in conversation with RoBERTa," *Comput. Res. Repository*, vol. 2108, p. 12009, Jan. 2021.

[41] A. Joshi, A. Bhat, A. Jain, A. Singh, and A. Modi, "COGMEN: Contextualized GNN based multimodal emotion recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2022, pp. 4148–4164.

Wei Ai received the Ph.D. degree from the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China, in 2016.

Her research interests include data mining, big data, cloud computing, and parallel computing.

Yuntao Shou is currently pursuing the bachelor's degree with the College of Computer Information and Engineering, Central South University of Forestry and Technology, Changsha, China.

His research interests include object detection and emotion recognition.

Tao Meng received the Ph.D. degree from the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China, in 2019.

His research interests include data mining, network analysis, and deep learning.

Keqin Li (Fellow, IEEE) is currently a SUNY Distinguished Professor of computer science with the State University of New York, New Paltz, NY, USA. He is also a National Distinguished Professor with Hunan University, Changsha, China. He has authored or coauthored over 880 journal articles, book chapters, and refereed conference papers. He holds nearly 70 patents announced or authorized by the Chinese National Intellectual Property Administration. His current research interests include cloud computing, fog computing and mobile edge computing, energy-efficient computing and communication, embedded systems and cyber-physical systems, heterogeneous computing systems, big data computing, high-performance computing, CPU–GPU hybrid and cooperative computing, computer architectures and systems, computer networking, machine learning, and intelligent and soft computing.

Dr. Li is an Asia-Pacific Artificial Intelligence Association (AAIA) Fellow. He is also a member of Academia Europaea (Academician of the Academy of Europe). He has received several best paper awards. He has served on the Editorial Boards for IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON CLOUD COMPUTING, IEEE TRANSACTIONS ON SERVICES COMPUTING, and IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING. He is an Associate Editor of the *ACM Computing Surveys* and the *CCF Transactions on High Performance Computing*.

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 03,2025 at 01:07:51 UTC from IEEE Xplore. Restrictions apply.