GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation

Jiang Li\textsuperscript{a,b,c}, Xiaoping Wang\textsuperscript{a,b,c,*}, Guoqing Lv\textsuperscript{a,b,c} and Zhigang Zeng\textsuperscript{a,b,c}

\textsuperscript{a}School of Artificial Intelligence and Automation, Huazhong University of Science and Technology (HUST), Wuhan 430074, China
\textsuperscript{b}Key Laboratory of Image Processing and Intelligent Control, Ministry of Education, Wuhan 430074, China
\textsuperscript{c}Hubei Key Laboratory of Brain-inspired Intelligent Systems, Wuhan 430074, China

\section*{ARTICLE INFO}

\textbf{Keywords:}
Multimodal machine learning
Graph neural networks
Emotion recognition in conversation
Multimodal fusion

\section*{ABSTRACT}

Multimodal machine learning is an emerging area of research, which has received a great deal of scholarly attention in recent years. Up to now, there are few studies on multimodal Emotion Recognition in Conversation (ERC). Since Graph Neural Networks (GNNs) possess the powerful capacity of relational modeling, they have an inherent advantage in the field of multimodal learning. GNNs leverage the graph constructed from multimodal data to perform intra- and inter-modal information interaction, which effectively facilitates the integration and complementation of multimodal data. In this work, we propose a novel Graph network based Multimodal Fusion Technique (GraphMFT) for emotion recognition in conversation. Multimodal data can be modeled as a graph, where each data object is regarded as a node, and both intra- and inter-modal dependencies existing between data objects can be regarded as edges. GraphMFT utilizes multiple improved graph attention networks to capture intra-modal contextual information and inter-modal complementary information. In addition, the proposed GraphMFT attempts to address the challenges of existing graph-based multimodal conversational emotion recognition models such as MMGCN. Empirical results on two public multimodal datasets reveal that our model outperforms the State-Of-The-Art (SOTA) approaches with the accuracy of 67.90\% and 61.30\%.

\section{1. Introduction}

As is known to all, the prerequisite for an empathic response is that we can accurately recognize the other party’s emotions. Recently, the applications of Human-Computer Interaction (HCI) have received widespread attention. The ability that machine can provide services naturally like a human is a pressing issue. Inspired by the empathy that human generates, machines should be endowed with the ability to detect emotions before providing services. With the synergistic development of artificial intelligence and cognitive science, Emotion Recognition in Conversation (ERC) is playing an outstanding role in numerous applications of HCI. Examples include conversational robot [1], opinion mining [2], medical diagnosis [3].

Most ERC approaches model interactions based on intra-modal context. HiGRU [4] leverages two improved Gated Recurrent Units (GRUs) to model word-level information and utterance-level context, respectively, and achieves outstanding performance on three dialogue emotion datasets. DialogueRNN [5] updates contextual information of the utterance and state of the participant through global GRU and party GRU, respectively, while emotion GRU is used for decoding emotion representation of the utterance. COSMIC [6] improves DialogueRNN and introduces external knowledge. Hu et al. [7] are motivated by Cognitive Theory of Emotion [8] to model contextual information flow from a cognitive reasoning perspective. However, these recurrence-based methods have difficulty capturing information of long-distance utterances and focus on considering the nearest ones to update state of query utterance. Since Kipf et al. [9] proposed Graph Convolutional Network (GCN), Graph Neural Networks (GNNs) have been frequently applied to computer vision, natural language processing, recommendation systems, etc. GNNs can easily capture long-distance contextual information in ERC tasks through the powerful capability of relational modeling. DialoguGCN [10] models each conversation as a graph, which addresses the problem of context propagation in the recurrence-based methods. Ishiwatari et al. [11] consider both the speaker dependency and the sequential information by combining graph neural networks and relational position encodings. KET [12] adopts a dynamic context-aware graph attention mechanism to capture external commonsense knowledge and promote the understanding of each word in the utterances. DAG-ERC [13] models long-distance contextual information flow through a directed acyclic graph for emotion detection. Unfortunately, the recurrence-based and graph-based approaches mentioned above either neglect multimodal information or fail to consider inter-modal interaction information.

Multimodal machine learning has risen in popularity in recent years. Multimodality is originally proposed to compensate for the deficiency that unimodality does not sufficiently reflect the reality state in some circumstances. Multimodal ERC is influenced by the fact that humans perceive the world from multisensory information. Figure 1 illustrates a multimodal dialogue scenario whose sources mainly include the visual, acoustic and textual modalities.
ConGCN [14] models both utterances and speakers as nodes, as well as both context dependencies and speaker dependencies as edges, for multi-participant conversational emotion detection. But ConGCN only takes into account textual features and acoustic features. MMGCN [15] considers both multimodal information and long-distance contextual information through graph-based model, and shows its superiority over the baseline models on two multimodal datasets. Nevertheless, MMGCN coarsely connects the utterance to all other intra-modal utterances, which brings additional noise. Although GNNs have a high capability of modeling relations, which demands that the input to the model is a homogeneous graph [16], MMGCN integrates the utterance nodes of three modalities into a single GNN simultaneously, which coarsely treats the complex heterogeneous graph as a homogeneous graph as the direct input of GNN. Furthermore, Multimodal fusion bears the problem of data heterogeneity [17], and the more types of modality there are, the more difficult it is to fuse them.

In order to tackle the mentioned dilemmas, we present a novel multimodal fusion method based on graph neural networks, namely Graph network based Multimodal Fusion Technique (GraphMFT), for conversational emotion detection in this paper. GraphMFT utilizes multiple improved Graph AtTention networks (GATs) to extract intra-modal contextual dependencies and inter-modal complementary dependencies, which can effectively facilitate the interaction of current utterance with intra-modal and inter-modal utterances. The proposed GraphMFT differs significantly from MMGCN, mainly including: (1) We construct three graphs, each containing two modalities. GraphMFT employs three improved GATs, each only fusing information from two modalities. So we mitigate the problem of data heterogeneity in multimodal fusion techniques by reducing the number of modalities in the network. (2) Instead of connecting the current node to all nodes within the modality, we connect it to variable context nodes, canceling out the potential for additional noise. (3) We adopt a spatial-based GNN as the basic backbone network, which requires explicit construction of edges, while MMGCN adopts a spectral-based GNN, which is not necessary to construct edges explicitly, but only a big adjacency matrix. (4) GraphMFT leverages the graph attention module to learn the weights of edges, while MMGCN utilizes the angular similarity to represent these weights. (5) We design a new fusion network, while MMGCN directly applies the off-the-shelf network structure, i.e., GCNI [18].

Our main contributions in this work can be summarized as follows:

1. A graph-based multimodal fusion model, GraphMFT, is proposed. GraphMFT can effectively capture intra-modal contextual information and inter-modal complementary information. Meanwhile, since there are few multimodal ERC models based on GNNs, we contribute a new fusion architecture for multimodal machine learning.

2. We model a conversation as three graphs, i.e., V-A graph, V-T graph, and A-T graph. Each graph contains information from only two modalities, which decreases the difficulty of multimodal fusion. The nodes of each graph are the utterances of each modality in the conversation, and the edges of each graph include both intra-modal dependencies and inter-modal dependencies.

3. We conduct extensive experiments on two public benchmark datasets, i.e., IEMOCAP and MELD datasets. Experimental results show that our model outperforms all baseline methods.

The remaining of this paper is organized as follows. In Section 3, the architecture we propose is presented with a focus on graph network based multimodal fusion techniques. In Section 4, the experimental preparation and experimental setup are described. We next report the experimental results with an in-depth discussion and analysis in Section 5. Besides, Section 2 and Section 6 are an introduction to related works and a summary of this paper, respectively.

2. Related works

2.1. Emotion recognition in conversation

Emotion Recognition in Conversation (ERC) aims to determine the emotion state of each utterance in a conversation and plays a crucial role in affective dialogue systems. How to model the conversational context and multimodal interaction is the core of this task. According to the number of input modality, existing ERC models can be divided into unimodal approaches and multimodal approaches.

**Unimodal Approaches** utilize single modal input information to judge the emotion of each utterance. Current unimodal-based approaches focus on utilizing textual modality for conversational emotion detection. KET [12] exploits hierarchical self-attention to interpret contextual utterances and utilizes a context-aware graph attention mechanism to dynamically learn external commonsense knowledge, which improves the performance of ERC task. DialogXL [19] addresses the issue that the pre-trained language models can’t be applied in the hierarchical structures of conversational utterances. COSMIC [6] utilizes...
commonsense including mental states, events, and causal relations to learn interactions between interlocutors, which addresses challenges such as emotion shift detection and context propagation. In order to model the intrinsic structure of a conversation, DAG-ERC [13] encodes the utterances with a directed acyclic graph in an attempt to combine the advantages of graph neural networks and recurrent neural networks. Other unimodal ERC methods include Hi-GRU [4], DialogueGCN [10], and DialogueCRN [7]. With the advancement of multimodal machine learning and the emergence of increasingly complex conversational sentiment recognition scenarios, the unimodal approaches offer limited performance gains.

**Multimodal Approaches** detect emotions in conversations through multimodal information. As the information of unimodality is biased and susceptible to interference from external noise factors, the research on ERC models based on multimodal fusion has received widespread attention for their merit of the complementarity between different modalities [20]. CMN [21] utilizes multimodal information by means of directly concatenating three modalities features, and adopts the Gated Recurrent Unit (GRU) to model contextual information. ICON [22] extracts multimodal conversation features and leverages global memories to model emotional influences of the self- and inter-speakers hierarchically, which generates contextual summaries to improve the performance of utterance-videos emotion recognition. ConGCN [14] is a GCN-based multimodal fusion model for ERC task. In this method, the node represents an utterance or a speaker, the edge between two utterances represents the context-sensitive dependence, and the edge between a speaker and its corresponding utterance represents the speaker-sensitive dependence. MMGCN [15] is also a GCN-based model for the multimodal ERC task, which can capture not only long-distance contextual information but also multimodal interactive information. Other multimodal ERC methods include MFN [23], BC-LSTM [24], and DialogueRNN [5]. These methods either do not capture the interaction information between modalities or ignore the heterogeneity of multimodal data. Additionally, there are few multimodal-based ERC models, and more relevant research is necessary to advance the field of ERC. Based on this, we propose a new GNN-based multimodal fusion method for conversational emotion detection. The proposed model is intended to minimize the problems mentioned above, as well as to facilitate the development of the multimodal learning field.

### 2.2. Graph convolutional networks

Graphs naturally exist in many practical applications, such as social analysis and fraud detection. The data is represented as graphs that can encode the structural information to model the relations among entities. However, the irregular structure of the graph data makes it difficult to define convolution and filtering on graphs. To this end, Graph Convolutional Networks (GCNs) are put forward to conduct convolutional operations on the graph. GCNs have been playing an essential role in numerous fields, such as natural language processing, computer vision, recommender system. GCNs can be mainly categorized into spectral models and spatial models. Spectral models are defined based on graph Fourier Transform, while spatial models are defined as the aggregations of neighboring nodes. The widely-used GCNs include 1stChebNet [9], GraphSAGE [25], GAT [26]. In this work, unlike MMGCN which employs spectral GCN, we leverage a spatial-based approach (i.e. GAT) as the base backbone network.

### 2.3. Multimodal machine learning

The goal of multimodal machine learning is to relate and process information from multiple modalities, which is extraordinarily potential in many fields and attaching increasingly more importance. It is composed of primarily five parts: representation, translation, alignment, fusion, and co-learning, where multimodal fusion is one of the most researched topics [27]. CM-BERT [28] proposes a cross-modal BERT approach to fine-tune the pre-trained BERT model through the interaction of text and audio modalities. Ephrat et al. [29] propose a deep network-based framework which combines visual and auditory signals to separate a single speech signal from mixture sounds such as ambient noise or other speakers. Multilogue-Net [30] presents an end to end Recurrent Neural Network (RNN) architecture to extract the contextual information of all modalities in the dialogue and capture interactions between the available modalities. Sahay et al. [31] adopt the inter-modal interactions of a segment while taking into account the sequence of segments in a video to model the inter-segment inter-modal interactive relationship. Kumar et al. [32] suggest that learning cross-modal interactions is instrumental in solving the problems confronting multimodal sentiment analysis, and propose a gating mechanism for multimodal interaction and fusion. GME-LSTM [33] fuses multimodal information at the word level and leverages the gated multimodal embedding to alleviate the difficulties of fusing noisy modalities, which can effectively model the multimodal sequences of speech and perform sentiment comprehension. BBFN [34] fuses relevance increment and separate difference increment of pairwise modalities, and trains simultaneously the two parts to simulate the combat between them. Graph neural networks possess tremendous ability to model relationships and extract interactive information between different modalities. However, to our knowledge, there are still relatively few existing GNN-based multimodal fusion models. Therefore, we aspire to contribute a new GNN-based architecture to advance the development of multimodal learning.

### 3. The proposed architecture

This section contains a detailed presentation of method proposed in this paper. The overall architecture of our approach is shown in Figure 2, including unimodal encoding, graph network based multimodal fusion technique, and emotion prediction. Here, graph network based multimodal fusion technique mainly consists of speaker embedding,
GraphMFT: A graph network based multimodal fusion technique

![Diagram](image_url)

**Figure 2:** The overall architecture of our approach, which consists of unimodal encoding, graph network based multimodal fusion technique, and emotion prediction. Here, the flow diagram of multi-graph based intra- and inter-modal interactions can be seen in Figure 4.

3.1. Problem definition

In the multimodal ERC system, each conversation contains \( m \) utterances \( u_1, u_2, \ldots, u_m \), and each utterance \( u_i \) has 3 modal expressions \( u_i^V, u_i^A, u_i^T \). Thus, there exist a total of \( 3 \times m \) utterance expressions. Here, V, A, and T denote visual, acoustic, and textual modalities, respectively; we refer to \( u_i^V, u_i^A, u_i^T \) as visual, acoustic, and textual utterances, respectively. Each utterance \( u_i \) (containing three modalities, i.e., \( u_i^V, u_i^A, u_i^T \)) corresponds to an emotion \( y_i \). Each conversation involves two or more speakers, so two utterances may be from the same speaker or from different speakers. Given utterance \( u_i \), the goal of multimodal ERC task is to predict emotion \( y_i \) corresponding to \( u_i \) based on intra- and inter-modal interaction learning. The multimodal emotion datasets used include IEMOCAP, which includes 6 emotion labels, and MELD, which includes 7 emotion labels. Referring to the approach of MMGCN [15], the features of visual, acoustic and text modalities are extracted employing DenseNet [35], OpenSmile toolkit [36] and TextCNN [37], respectively.

3.2. Unimodal encoding

We exploit fully connected networks to encode utterances in the visual and acoustic modalities separately to enhance the expressiveness of utterance features. For textual modal utterances, a bidirectional Long Short-Term Memory (LSTM) network is leveraged to extract the contextual information of current utterance. The process of unimodal encoding can be described by the following equations:

\[
\begin{align*}
    x_i^V &= W_i^V c_i^V + b_i^V, \\
    x_i^A &= W_i^A c_i^A + b_i^A, \\
    x_i^T &= \text{LSTM}_v(c_i^T, x_{i-1}^T, x_{i+1}^T),
\end{align*}
\]

where \( c_i^V, c_i^A, \) and \( c_i^T \) are inputs of unimodal encoding; \( x_i^V, x_i^A, x_i^T \) are outputs of unimodal encoding; V, A, and T denote visual, acoustic, and textual modalities, respectively; \( \text{LSTM}_v \) denotes the bidirectional LSTM network; \( W_i^V, W_i^A, b_i^V, b_i^A \) are the trainable parameters.

3.3. Graph network based multimodal fusion technique

In order to encode both intra-modal contextual information and inter-modal complementary information, we propose a Graph network based Multimodal Fusion Technique (GraphMFT) for Emotion Detection in Conversation. The proposed GraphMFT allows an utterance of current modality to interact with both intra-modal utterances and inter-modal utterances at the same time.

3.3.1. Speaker embedding

Referring to Hu et al. [15], we encode speakers as embedding vectors to capture the characteristics of speakers. The speaker embedding is then summed with utterance feature of each modality to obtain new utterance feature carrying speaker information. The above process can be represented as follows:

\[
S_{pk} = \text{Embedding}(\text{SPK}, s), \quad X^\tau = X^\tau + S_{pk}, \quad \tau \in \{V, A, T\},
\]

where \( s \) is the number of speakers, and \( \text{SPK} \) is the speaker array (note that different speakers are represented as different index values); Embedding denotes the embedding layer, and \( S_{pk} \) is the speaker embedding obtained by feeding the speaker array and the number of speakers to the embedding layer; \( X^\tau \) is the feature set after unimodal encoding (\( x_i^\tau \in X^\tau \)); \( X^\tau \) denotes the feature set adding speaker information, and V, A, T indicate visual, acoustic, textual modalities, respectively.

3.3.2. Graph construction

We construct three graphs for each dialogue, including V-A graph, V-T graph, and A-T graph. These graphs can be represented as \( G_{VA} = (V^VA, e^VA, W^VA), G_{VT} = \)

\[
\]
Improved GAT; improved GAT be formalized as:

\[
X^{(l)} = X^{(l-1)} + \text{MultiGAT}(X^{(l-1)}, \mathcal{E}),
\]

\[
H = W_{lm}[X^{(0)} \| X^{(1)} \| \cdots \| X^{(L)}],
\]

where \(l = 1, \cdots, L\), and \(L\) is the number of network layers; MultiGAT denotes the multi-head graph attention layer; \(X^{(l-1)}\) is the \((l-1)\)-th layer feature matrix, and \(\mathcal{E}\) is the edge set, and they are both inputs to GAT; \(H\) is the output of GAT, \(W_{lm}\) is the trainable parameter.

We employ three GATs in order to extract both intra- and inter-modal information, where the input of each GAT includes information from two modalities. As shown in Figure 4, we take V-A graph, V-T graph, and A-T graph as the inputs of improved GATs; then the feature matrices of intra- and inter-modal information interactions are obtained. Taking node \(u^V_i\) in V-A graph as an example, \(u^V_i\) collects both visual contextual information (i.e., the information of \(u^V_{i-p}, \cdots, u^V_{i-1}, u^V_{i+1}, \cdots, u^V_{i+p}\)) and acoustic complementary information (i.e., \(u^A_i\)) through GATs. We can describe computational procedure of V-A graph as:

\[
X^{VA,(l)} = X^{VA,(l-1)} + \text{MultiGAT}(X^{VA,(l-1)}, \mathcal{E}^{VA}),
\]

\[
H^{VA} = W_{lm}[X^{VA,(0)} \| X^{VA,(1)} \| \cdots \| X^{VA,(L)}],
\]

where \(l = 1, \cdots, L\), and \(L\) is the number of layers of improved GAT; \(X^{VA,(l)} = X^V(l) \cup X^A(l)\), and \(X^{VA,l}\) is...
the $l$th layer node feature matrix in $V$-A graph; $X^{VA,(0)} = X^V \cup X^A$ is the input of improved GAT, $H^{VA}$ is the output of improved GAT; $e^{VA}$ is the edge set in $V$-A graph, and MultiGAT is multi-head graph attention layer; $\parallel$ denotes the concatenation operation, and $W_{va}$ is the learnable parameter.

Similarly, according to the computational procedure of $V$-A graph, $V$-T graph and $A$-T graph are calculated as:

\[
\begin{align*}
X^{VT,(l)} &= X^{VT,(l-1)} + \text{MultiGAT}(X^{VT,(l-1)}, E^{VT}), \\
H^{VT} &= W_{vt}[X^{VT,(0)} \parallel X^{VT,(1)} \parallel \cdots \parallel X^{VT,(L)}], \\
X^{AT,(l)} &= X^{AT,(l-1)} + \text{MultiGAT}(X^{AT,(l-1)}, E^{AT}), \\
H^{AT} &= W_{at}[X^{AT,(0)} \parallel X^{AT,(1)} \parallel \cdots \parallel X^{AT,(L)}],
\end{align*}
\]

where $H^{VT}$ and $H^{AT}$ are the outputs of two improved GATs; $W_{vt}$ and $W_{at}$ are the learnable parameters. Then, we sum feature matrices of the same modalities,

\[
\begin{align*}
H^V &= H^V_a + H^V_t, \\
H^A &= H^A_a + H^A_t, \\
H^T &= H^T_v + H^T_a,
\end{align*}
\]

where $H^V_a, H^A_a \subseteq H^{VA}$ (moreover, $H^{VA} = H^V_a \cup H^V_t$, $H^A_a$ and $H^A_t$ are the node feature matrices updated by improved GAT in $V$-A graph; the rest can be deduced by analogy, $H^T_v, H^T_a \subseteq H^{VT}$, and $H^T_v, H^T_a \subseteq H^{AT}$; $H^V, H^A$, and $H^T$ are the visual, acoustic, and textual feature matrices through multimodal fusion, respectively.

### 3.4. Emotion prediction

After multimodal fusion, we apply the concatenation operation to $H^V, H^A$, and $H^T$. Finally, we obtain the final feature matrix $H$ carrying the three modal information. $H$ is computed by the following:

\[
H = W_{vat}[H^V \parallel H^A \parallel H^T],
\]

where $\parallel$ is the concatenation operation and $W_{vat}$ is the learnable parameter. We take $h_i \in H$ as the input to a fully connected network for emotion prediction:

\[
\begin{align*}
\hat{y}_i &= \text{ReLU}(W_c h_i + b_c), \\
p_i &= \text{Softmax}(W'_c \hat{y}_i + b'_c), \\
\hat{y}_i &= \text{argmax}(p_i[k]),
\end{align*}
\]

where $h_i \in H$ is the final feature vector of the $i$th utterance $u_i$, and ReLU denotes the non-linear activation function; $p_i$ denotes the probability distribution of a predicted emotion of $u_i$, and Softmax denotes the softmax function; $\hat{y}_i$ denotes the predicted emotion; $W_c, W'_c, b_c$, and $b'_c$ are the trainable parameters.

### 3.5. Objective training

We employ cross-entropy loss as objective function to train the proposed model, which can be formalized as:

\[
\text{Loss} = -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{j=0}^{n(i)-1} y_{ij} \log p_{ij} + \lambda \|W_{is}\|_2^2,
\]

where $N$ is the number of all conversations in the dataset, and $n(i)$ is the number of utterances of the $i$-th conversation; $y_{ij}$ denotes the ground-truth emotion of the $j$-th utterance in the $i$-th conversation, and $p_{ij}$ denotes the probability distribution of predicted emotion of the $j$-th utterance in the $i$-th conversation; $\lambda$ is the L2-regularizer weight, and $W_{is}$ is the set of all learnable parameters.

### 4. Experiments

This section focuses on several experiment-related contents, including the baseline methods for comparison, the public benchmark datasets, and the detailed experimental settings.

#### 4.1. Baseline methods

To demonstrate the effectiveness of GraphMFT in multimodal modeling, we compare it with some baseline models. The baseline methods we employ are as follows. TFN [39] allows dynamic modeling of intra- and inter-modality in an end-to-end manner. In this model, intra-modality is modeled by three modal embedding subnetworks, and inter-modality is modeled by a method called tensor fusion. MFN [23] fuses multi-views information and unifies the multimodal features, but it ignores context-aware dependencies and speaker-aware dependencies. BC-LSTM [24] exploits textual, visual, and acoustic modalities to recognize emotion, and it captures multimodal information and extracts context-dependent features through a contextual LSTM network. DialogueCNN [5] tracks the individual party states in the conversation for emotion classification, where the characteristic of the speaker is taken into account for each incoming utterance to provide it with a more fine-grained context. DialogueGCN [10] is a graph-based model, where the conversation is regarded as a graph, which can efficiently capture long-range contextual information. We directly concatenate multimodal features to implement multimodal setting. DialogueCRN [7] investigates cognitive factors for ERC, where the cognitive phase is introduced to extract and fuse emotional clues from context and emotional clues are successfully utilized to better classify emotion states. The multimodal features are concatenated to achieve multimodal setting. MMGCN [15] designs a graph which can model the intra-speaker context and inter-modal dependencies to utilize multi-modal information as well as long-distance contextual information.

### 4.2. Datasets

We conducted extensive experiments on two widely used public datasets, MELD [40] and IEMOCAP [41]. They are both multimodal benchmark datasets containing visual, acoustic and textual modalities. The statistics of the two datasets adopted are shown in Table 1.

MELD dataset is a multi-party and multimodal dataset for ERC task. It includes textual, visual and acoustic modalities information and over 1400 conversations, a total of 13700 utterances, which are categorized to seven emotions: Neutral, Surprise, Fear, Sadness, Joy, Disgust, and Angry.
Table 1
The statistics of two public benchmark datasets adopted. The left shows the number of conversations, utterances, and classes. The right shows the distribution of sample size for each emotion.

|                | *Conversations | *Utterances | *Classes | *Neutral | *Surprise | *Fear | *Sadness | *Joy | *Disgust | *Angry |
|----------------|----------------|-------------|----------|----------|-----------|-------|----------|------|----------|--------|
|                | train | valid | test   | train | valid | test   |         |        |          |        |
| MELD           | 1039  | 114   | 280    | 9989   | 1109   | 2610   | 7   | 6436    | 1636   | 358     | 1002   |
| IEMOCAP        | 120   | 31    |        | 5810   | 1623   |        | 6   | 648     | 1084   | 1708    | 1103   |

*Joy, *Surprise, *Sad, 1e-05, *Excited, *Neutral, *Frustrated

Table 2
Description of the partial hyperparameter settings.

| Dataset       | Number of layers | Learning rate | Batch size | Regularization factor | Dropout rate |
|---------------|------------------|---------------|------------|-----------------------|--------------|
| MELD          | 3                | 2e-05         | 32         | 1e-05                 | 0.5          |
| IEMOCAP       | 5                | 1e-05         | 16         | 1e-05                 | 0.5          |

It has three or more speakers in each conversation and is collected from the TV show Friends.

IEMOCAP dataset is an ERC dataset containing multimodal audio-video-text conversations from ten professional actors. It contains 151 conversations, a total of 7433 dyadic utterances, where each utterance corresponds to an emotion label. The emotion labels of these utterances are divided into six categories, including Happy, Sad, Neutral, Angry, Excited, and Frustrated.

It should be noted that the MELD dataset differs from the IEMOCAP dataset in that MELD suffers from a severe class-imbalanced problem. The right of Table 1 shows the distribution of sample size for each class in both datasets. It can be obviously seen that the sample size of Fear and Disgust in MELD is much smaller than that of other emotions, and such class-imbalanced problem poses a great challenge to emotion classification.

4.3. Detailed settings

We perform all experiments on NVIDIA GeForce RTX 3090, and the operating system is Ubuntu 20.04. Pytorch is used as deep learning framework and its version is 1.12.0. All experiments are implemented with Python 3.9.12. We choose AdamW as the optimizer to train proposed model. The maximum epoch is set to 100, the L2 regularization factor and dropout rate are set to 0.00001 and 0.5, respectively. For MELD dataset, the number of network layers is 3; the learning rate is 0.00002; the batch size is 32. For IEMOCAP dataset, the number of network layers is 5; the learning rate is 0.00001; the batch size is 16. Table 2 describes the partial hyperparameter settings. To evaluate the superiority of our proposed GraphMFT, we apply weighted-average F1 score and accuracy score as evaluation metrics.

5. Results and analysis

We report experimental results to evaluate the proposed GraphMFT in this section. First, GraphMFT is compared with all baseline methods as a whole. Then, the influence of different settings on our GraphMFT is discussed. Finally, we also carry out limitations analysis of our approach.

5.1. Overall results

The experimental results of all models are recorded in Table 3. It can be significantly seen that our GraphMFT outperforms all baseline models. On the IEMOCAP dataset, the accuracy score and weighted-average F1 score of GraphMFT are 67.90% and 68.07%, respectively, which are distinctly higher than those of SOTA method (i.e., MMGCN). DialogueCRN is another strong baseline model in addition to MMGCN. From Table 3, we can conclude that F1 score of GraphMFT on the IEMOCAP dataset is 2.73% higher than that of DialogueCRN. We record F1 score for each emotion on the IEMOCAP dataset. We found that, except for Happy and Frustrated, GraphMFT’s F1 scores for other emotions have significant improvement compared to other models. Notably, GraphMFT performs better for Sad and Excited than for other emotions, achieving F1 scores of 83.12% and 76.92%, respectively. In addition, the accuracy score of GraphMFT on the MELD dataset is improved by 1.99% relative to that of MMGCN. In general, the proposed GraphMFT shows superior performance than MMGCN and DialogueCRN in multimodal fusion.

5.2. Comparison of vanilla and improved GATs

To demonstrate the validity of the improved GAT, we replace the improved GAT module in GraphMFT with the vanilla GAT. We compare the experimental results with vanilla GAT and improved GAT on these two datasets, as displayed in Table 4. It is observed that our GraphMFT with improved GAT module achieves a 1.60% improvement of weighted-average F1 score on the IEMOCAP dataset in comparison to the model with vanilla GAT. The similar results appear on the MELD dataset. For instance, the model using vanilla GAT module achieves an accuracy score of 60.19% on the MELD dataset, which is 1.11% lower than GraphMFT using improved GAT. Therefore, our improved GAT can be an effective module to promote the performance of emotion classification.

5.3. Impact of number of network layers

When the number of network layers reaches to a specific threshold, the performance of GNNs drops dramatically if the number of network layers is increased further. This is notorious over-smoothing problem in graph models [9]. We make an improvement to standard GAT in Section 3.3.3 based on the idea of ResGCN. In order to exhibit that the
GraphMFT: A graph network based multimodal fusion technique

Table 3
The overall results of all models. Evaluation metrics contain Acc, F1, and wa-F1, which denote accuracy score (%), F1 score (%), and weighted-average F1 score (%), respectively. Bold font indicates the best performance.

| Model   | IEMOCAP          | MELD          |
|---------|------------------|---------------|
|         | Happy | Sad | Neutral | Angry | Excited | Frustrated |         | Acc | wa-F1 |         | Acc | wa-F1 |
| TFN     | 38.29 | 64.12 | 51.43 | 55.34 | 57.95 | 57.58 | 56.12 | 55.54 |
| BC-LSTM | 32.63 | 70.34 | 51.14 | 63.44 | 67.91 | 61.06 | 59.58 | 59.10 |
| MFN     | 47.19 | 72.49 | 55.38 | 63.04 | 64.52 | 61.91 | 60.14 | 60.32 |
| DialogueRNN | 33.18 | 78.80 | 59.21 | 65.28 | 71.86 | 58.91 | 63.40 | 62.75 |
| DialogueGCN | 51.59 | 74.54 | 62.38 | 67.25 | 73.96 | 59.97 | 65.31 | 65.34 |
| DialogueCRN | 45.45 | 77.53 | 61.99 | 66.67 | 72.04 | 64.12 | 65.56 | 65.71 |
| MMGCN   | 45.99 | 83.12 | 63.08 | 70.30 | 76.92 | 63.84 | 67.90 | 68.07 |

Table 4
The experimental results with vanilla GAT and improved GAT.

| Module   | IEMOCAP | MELD |
|----------|---------|------|
| Vanilla GAT | 66.97  | 66.47 | 60.19 | 57.90 |
| Improved GAT | 67.90 | 68.07 | 61.30 | 58.37 |

proposed approach alleviates the over-smoothing problem, we replace the improved GAT module in GraphMFT with the vanilla GAT, and provide a comparison of the experimental results using the vanilla GAT and the improved GAT. Figure 5 shows the effect of the number of network layers on the model performance. Note that the reported experimental results are based on the ICMOCAP dataset.

Figure 5: Experimental results for testing different number of network layers on the IEMOCAP dataset.

We can visualize that the performance of the model replaced with the vanilla GAT module decreases sharply as the number of network layers increases. When a certain threshold is reached, the performance of the model fluctuates around a lower value as the number of network layers is varied. Such result prohibits the possibility to improve the expressiveness of the model through increasing the number of network layers. As we expected, the performance of GraphMFT utilizing the improved GAT increases with the number of network layers in the beginning stage; when the number of network layers reaches a certain threshold, the performance of GraphMFT slowly decreases as the number of layers increases. Distinct from the vanilla GAT module, GraphMFT with improved GAT has a very weak performance degradation.

5.4. Impact of speaker embedding

Previous baseline approaches, such as MMGCN, prove that the characteristics of speakers play an essential part in ERC tasks. Table 5 shows the effect of adding or not adding speaker embeddings on the performance of GraphMFT. w/o (w) indicates without (with) speaker embeddings.

Table 5
The effect of adding or not adding speaker embeddings on the performance of GraphMFT. w/o (w) indicates without (with) speaker embeddings.

| Speaker Embedding | IEMOCAP | MELD |
|-------------------|---------|------|
|                   | Acc | wa-F1 | Acc | wa-F1 |
| w/o               | 65.74 | 65.61 | 61.03 | 58.04 |
| w                 | 67.90 | 68.07 | 61.30 | 58.37 |

Therefore, we can conclude that adding speaker features can improve the performance of ERC task. In addition, we realize that GraphMFT without speaker information and MMGCN with speaker information have comparable performance, which indicates that the proposed GraphMFT has excellent multimodal modeling capability.
5.5. Comparison of different modal settings

To investigate the effect of different modal settings on the performance of ERC model, we report in Table 6 the experimental results for the model with these settings. We can clearly see that the performance of three-modality-based models outperforms that of all two-modality-based models, indicating that more modalities contribute to the accuracy of emotion recognition. For the two-modal settings, the performance of the model based on visual and acoustic modalities is the worst, while the model with a combination of acoustic and textual modalities performs the best. Therefore, it can be concluded that the acoustic and textual modalities have more contributions to GraphMFT compared to the visual modality.

5.6. Impact of number of contextual nodes

When we create intra-modal edges, current node needs to connect the past $P$ and the future $F$ contextual nodes. This section investigates the impact of various $(P, F)$ combinations on the performance of our model. Specifically, we set the following $(P, F)$ combinations: (a) for the IEMOCAP dataset, the combinations are set to $(4, 4), (8, 8), \ldots, (40, 40)$; (b) for the MELD dataset, the combinations are set to $(1, 1), (2, 2), \ldots, (10, 10)$. Figure 6 depicts the effect of different $(P, F)$ combinations on accuracy scores and weighted-average F1 scores of GraphMFT. It can be concluded from Figure 6a that the performance of GraphMFT on the IEMOCAP dataset increases with increasing $P$ or $F$; when reaching a certain threshold ($16, 16$), both accuracy score and F1 score of GraphMFT start to fluctuate up and down. Figure 6b shows the performance variation of proposed model on the MELD dataset. What can be expected is that the variation pattern of the performance on the MELD dataset and the IEMOCAP dataset are nearly identical. Another interesting phenomenon is that GraphMFT requires 16 contextual nodes for optimal performance on the IEMOCAP dataset, while only 5 contextual nodes are required on the MELD dataset. The probable reason is that some neighboring utterances in the MELD dataset are not contiguous in the actual scenario, and thus the long-range context modeling is not required.

5.7. Limitations

As shown in Figure 7, we discuss the limitations of current ERC models by leveraging two confusion matrices. One of the challenges of ERC task is the Similar Emotion problem, meaning that many models have difficulty distinguishing similar emotions. Like existing models, our GraphMFT sometimes misclassifies similar emotions. As shown in Figure 7a, GraphMFT incorrectly detects ground-truth emotion Angry as Frustrated with a 26.47% probability; the ground-truth emotion Excited is incorrectly detected as Happy with a 12.71% probability. Another annoying problem in ERC task is that models are eager to recognize many emotions as Neutral, which we refer to as the non-Neutral Emotion problem. Most existing ERC models rely on text modality. Thus, when some non-Neutral utterances such as "Okay." and "Yeah." appear in the conversation, the model may detect them as Neutral directly. As shown in Figure 7a, the ground-truth emotion Happy is identified as Neutral with a probability of 20.14%.

Fortunately, the proposed GraphMFT can alleviate Similar Emotion problem compared to the baseline models such as MMGCN. By comparing Figure 7a and Figure 7b, it can be concluded that GraphMFT reduces the probability that Angry is identified as Frustrated from 29.41% of MMGCN to 26.47%; similarly, the probability that Excited is identified as Happy is reduced from 14.38% of MMGCN to 12.71% of GraphMFT. In addition, GraphMFT also mitigates non-Neutral Emotion problem. For example, the probability that Sad is classified as Neutral is reduced by 3.67% from 10.61% of MMGCN; GraphMFT reduces the probability that Excited is classified as Neutral from 17.39% to 10.70% of MMGCN. Hence, GraphMFT can compensate textual modal misinformation by visual and acoustic modal information, and adequately capture multimodal complementary information to improve the ability of emotion recognition.

6. Summary

To overcome the challenges of existing graph-based multimodal ERC models, we propose a novel multimodal fusion method, namely Graph network based Multimodal Fusion Technique (GraphMFT). The proposed GraphMFT models each conversation as three heterogeneous graphs, i.e., V-A graph, V-T graph, and A-T graph. Each graph includes data from only two modalities to diminish the difficulty of multimodal fusion. In addition, to handle the over-smoothing problem of graph neural networks, we improve the vanilla graph attention network inspired by ResGCN. We evaluate our proposed GraphMFT on two benchmark datasets that are widely used by most ERC models. Experimental results demonstrate that our proposed model can effectively perform intra- and inter-modal interactions to extract contextual information and complementary information. By comparing with the previous baseline model, GraphMFT prevails with a significant superiority.

Future studies could fruitfully and continuously explore the application of multimodal machine learning for conversational emotion recognition. Meanwhile, with the relational modeling capability of graph neural networks, we hope to further extend the graph-based models to the field of multimodal dialogue generation.
GraphMFT: A graph network based multimodal fusion technique

Figure 6: The effect of different combinations ($P, F$) on the performance of our model. Here, $P$ (or $F$) denotes the number of past (or future) contextual nodes of current node.

Figure 7: The confusion matrices of proposed GraphMFT and previous MMGCN on the IEMOCAP dataset. GraphMFT, like MMGCN, faces the problems of Similar-Emotion and non-Neutral emotion.

CRediT authorship contribution statement

**Jiang Li:** Conceptualization, Methodology, Data Curation, Software, Validation, Formal analysis, Investigation, Visualization, Writing - original draft, Project Administration, Writing - review & editing. **Xiaoping Wang:** Supervision, Writing - review & editing, Funding Acquisition. **Guoqing Lv:** Investigation, Writing - review & editing. **Zhigang Zeng:** Supervision, Funding Acquisition.

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.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62236005, 61936004, and U1913602.

Data availability

Data will be made available on request.
References

[1] A. M. Sabelli, T. Kanda, N. Hagitu, A conversational robot in an elderly care center: An ethnographic study, in: Proceedings of the International Conference on Human-Robot Interaction, IEEE, 2011, pp. 37–44. doi:10.1109/1957656.1957669.

[2] K. Cortis, B. Davis, Over a decade of social opinion mining: A systematic review, Artificial Intelligence Review 54 (7) (2021) 4873–4965. doi:10.1007/s10462-021-10038-2.

[3] A. Bhavan, P. Chauhan, Hitkul, R. R. Shah, Bagged support vector machines for emotion recognition from speech, Knowledge-Based Systems 184 (2019) 1–7. doi:10.1016/j.knosys.2019.104886.

[4] W. Jiao, H. Yang, I. King, M. R. Lyu, Higru: Hierarchical gated recurrent units for utterance-level emotion recognition, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 397–406. doi:10.18653/v1/N19-1037.

[5] 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 and Thirty-First Innovative Applications of Artificial Intelligence Conference and AAAI Symposium on Educational Advances in Artificial Intelligence, Vol. 33, AAAI Press, 2019, pp. 6818–6825. doi:10.1609/aaai.v33i01.33016818.

[6] W. Shen, S. Wu, Y. Yang, X. Quan, Directed acyclic graph network for conversational emotion recognition, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2020, pp. 1551–1560. doi:10.18653/v1/2020.acl-long.123.

[7] D. Hu, L. Wei, X. Huai, Dialoguercn: Contextual reasoning networks for emotion recognition in conversations, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2021, pp. 7042–7052. doi:10.18653/v1/2021.acl-long.547.

[8] R. S. Lazarus, Progress on a cognitive-motivational-relational theory of emotion, American Psychologist 46 (8) (1991) 819. doi:10.1037/0003-066X.46.8.819.

[9] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, 2017. URL https://openreview.net/forum?id=SJUkaygYl

[10] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, Dialoguercn: A graph convolutional neural network for emotion recognition in conversations, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2019, pp. 154–164. doi:10.18653/v1/D19-1015.

[11] T. Ishiwatari, Y. Yasuda, T. Miyazaki, J. Goto, Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020, pp. 7360–7370. doi:10.18653/v1/2020.emnlp-main.597.

[12] P. Zhong, D. Wang, C. Miao, Knowledge-enriched transformer for emotion detection in textual conversations, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2019, pp. 165–176. doi:10.18653/v1/D19-1016.

[13] W. Shen, S. Wu, Y. Yang, X. Quan, Directed acyclic graph network for conversational emotion recognition, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2021, pp. 1551–1560. doi:10.18653/v1/2021.acl-long.123.

[14] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, G. Zhou, Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations, in: Proceedings of the International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 5415–5421. doi:10.24963/ijcai.2019/752.

[15] 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 Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2021, pp. 5666–5675. doi:10.18653/v1/2021.acl-long.446.

[16] W. Su, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, A comprehensive survey on graph neural networks, IEEE Transactions on Neural Networks and Learning Systems 32 (1) (2021) 4–24. doi:10.1109/TNNLS.2020.2978386.

[17] Z. Jia, Y. Lin, J. Wang, Z. Feng, X. Xie, C. Chen, Hetemotionnet: two-stream heterogeneous graph recurrent neural network for multimodal emotion recognition, in: Proceedings of the ACM International Conference on Multimedia, Association for Computing Machinery, 2021, pp. 1047–1056. doi:10.1145/3474885.3475583.

[18] M. Chen, Z. Wei, Z. Huang, B. Ding, Y. Li, Simple and deep graph convolutional networks, in: Proceedings of the International Conference on Machine Learning, PMLR, 2020, pp. 1725–1735. doi:10.5555/3394838.3394999.

[19] W. Shen, J. Chen, X. Quan, Z. Xie, Dialoggl: All-in-one xlnet for multi-party conversation emotion recognition, Vol. 35, 2021, pp. 13789–13797. doi:10.1609/aaai.v35i15.17625.

[20] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Information Fusion 37 (2017) 98–125. doi:10.1016/j.inffus.2017.02.003.

[21] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2122–2132. doi:10.18653/v1/N18-1193.

[22] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, Icon: Interactive conversational memory network for multimodal emotion detection, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2594–2604. doi:10.18653/v1/D18-1280.

[23] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, L.-P. Morency, Memory fusion network for multi-view sequential learning, Vol. 32, 2018, pp. 5634–5641. doi:10.1609/aaai.v32i1.12021.

[24] 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 Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2017, pp. 873–883. doi:10.18653/v1/P17-1081.

[25] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, Advances in neural information processing systems 30 (2017). doi:10.18653/v1/2021.acl-long.123.

[26] P. Vešličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in: International Conference on Learning Representations, 2018, pp. 1–12. URL https://openreview.net/forum?id=JXZMspkCZ

[27] T. Baltrusaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2) (2019) 423–443. doi:10.1109/TPAMI.2018.2798607.

[28] K. Yang, H. Xu, K. Gao, Cm-bert: Cross-modal bert for text-audio sentiment analysis, in: Proceedings of the ACM International Conference on Multimedia, Association for Computing Machinery, 2020, p. 521–528. doi:10.1145/3394711.3431690.
GraphMFT: A graph network based multimodal fusion technique

[29] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, M. Rubinstein, Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation, ACM Transactions on Graphics 37 (4) (2018) 1–11. doi:10.1145/3197517.3201357.

[30] A. Shenoy, A. Sardana, Multilogue-net: A context-aware rnn for multi-modal emotion detection and sentiment analysis in conversation, in: Second Grand-Challenge and Workshop on Multimodal Language, Association for Computational Linguistics, 2020, pp. 19–28. doi:10.18653/v1/2020.chALLENGEhml-1.3.

[31] S. Sahay, S. H. Kumar, R. Xia, J. Huang, L. Nachman, Multimodal relational tensor network for sentiment and emotion classification, in: Proceedings of Grand Challenge and Workshop on Human Multimodal Language, Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), Association for Computational Linguistics, 2018, pp. 20–27. doi:10.18653/v1/W18-3303.

[32] A. Kumar, J. Vepa, Gated mechanism for attention based multi modal sentiment analysis, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2020, pp. 4477–4481. doi:10.1109/ICASSP40776.2020.9053012.

[33] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, L.-P. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: Proceedings of the ACM international conference on multimodal interaction, Association for Computing Machinery, 2017, p. 163–171. doi:10.1145/3136755.3136801.

[34] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-P. Morency, S. Poria, Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis, in: Proceedings of the International Conference on Multimodal Interaction, Association for Computing Machinery, 2021, p. 6–15. doi:10.1145/3462244.3479919.

[35] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.

[36] B. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge, Speech Communication 53 (9–10) (2011) 1062–1087. doi:10.1016/j.specom.2011.01.011.

[37] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 2014, pp. 1746–1751. doi:10.3115/v1/D14-1181.

[38] G. Li, M. Miller, G. Qian, I. C. D. Perez, A. Abualshour, A. K. Thabet, B. Ghanem, Deepgcns: Making gcns go as deep as cnns, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021). doi:10.1109/TPAMI.2021.3074057.

[39] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 1103–1114. doi:10.18653/v1/D17-1115.

[40] 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 Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 527–536. doi:10.18653/v1/p19-1858.

[41] 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, Language Resources and Evaluation 42 (4) (2008) 335–359. doi:10.1007/s10579-008-9076-6.