2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.01036

Multivariate, Multi-frequency and Multimodal: Rethinking Graph Neural Networks for Emotion Recognition in Conversation

Feiyu Chen[†‡] Jie Shao[†‡*] Shuyuan Zhu[†] Heng Tao Shen[†‡]
[†]University of Electronic Science and Technology of China, Chengdu, China
[‡]Sichuan Artificial Intelligence Research Institute, Yibin, China
{chenfeiyu,shaojie,eezsy,shenhengtao}@uestc.edu.cn

Abstract

Complex relationships of high arity across modality and context dimensions is a critical challenge in the Emotion Recognition in Conversation (ERC) task. Yet, previous works tend to encode multimodal and contextual relationships in a loosely-coupled manner, which may harm relationship modelling. Recently, Graph Neural Networks (GNN) which show advantages in capturing data relations, offer a new solution for ERC. However, existing GNN-based ERC models fail to address some general limits of GNNs, including assuming pairwise formulation and erasing high-frequency signals, which may be trivial for many applications but crucial for the ERC task. In this paper, we propose a GNN-based model that explores multivariate relationships and captures the varying importance of emotion discrepancy and commonality by valuing multi-frequency signals. We empower GNNs to better capture the inherent relationships among utterances and deliver more sufficient multimodal and contextual modelling. Experimental results show that our proposed method outperforms previous state-of-the-art works on two popular multimodal ERC datasets.

1. Introduction

Human beings constantly express their feelings in everyday communication. Emotion Recognition in Conversation (ERC) aims at enabling machines to detect interactive human emotions in a dialogue, utilizing multi-sensory data, including textual, visual and acoustic information [5, 13, 18, 24]. Unlike traditional affective computing tasks that are performed on single modalities (e.g., text, speech or facial images) [12, 28, 32] or/and in non-conversational scenario [15, 23, 33], there exists a distinct and essential challenge in the ERC task - the complex multivariate relationships among multiple modalities and conversational context. In other words, the emotional dependencies of an utterance are usually of high arity, and involve multi-source information across both modality and context dimensions. Figure 1 presents a sample of conversation between two speakers. Take the utterance u_3 as an example. The visual and acoustic messages of utterance u_3 (an expressionless face and a flat tone) are ambiguous, but imply a veiled anger if coupled with the text. Moreover, the emotion behind u_3 is also related to the preceding context u_1 and u_2. In particular, the change from calling by nickname in u_1 to calling by full name in u_3 suggests an emotion shift caused by u_2, since another speaker tries to make a joke with a pretended lightness. Therefore, the relationships in $\{u_1, u_2, u_3\}$ are complex and multivariate, and involve interdependencies across both modality and context dimensions.

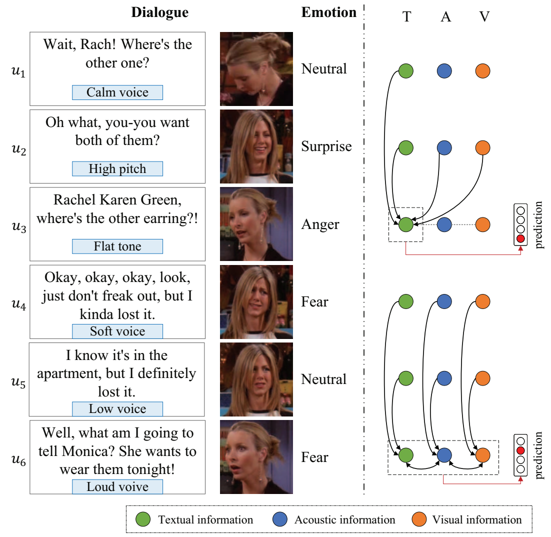

Dialogue
Emotion
T
A
V
u_1
Wait, Rach! Where's the other one?
Calm voice
Neutral
u_2
Oh what, you-you want both of them?
High pitch
Surprise
u_3
Rachel Karen Green, where's the other earring?!
Flat tone
Anger
prediction
u_4
Okay, okay, okay, look, just don't freak out, but I kinda lost it.
Soft voice
Fear
u_5
I know it's in the apartment, but I definitely lost it.
Low voice
Neutral
u_6
Well, what am I going to tell Monica? She wants to wear them tonight!
Loud voive
Fear
prediction
Textual information
Acoustic information
Visual information

Figure 1. An example of multimodal dialogue (left) and the complex multivariate relationships of u_3 and u_6 (right).

*Corresponding author: Jie Shao. This work is supported by the National Natural Science Foundation of China (No. 61832001 and No. 62276047), Natural Science Foundation of Sichuan Province (No. 2023NSFSC1972) and Science and Technology Program of Yibin Sanjiang New Area (No. 2023SJXQYBKJJH001).

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.01036

Researchers have been exploring how to capture the complex relationships more effectively. Among existing ERC models, a dominant paradigm is to capture contextual relationships with context-sensitive modules such as recurrent unit or transformer, whilst modelling multimodal relationships through various fusion methods [4, 24, 25, 34]. Despite the advances, this paradigm tends to underrate multivariate relationships among modalities and context, as it limits the natural interaction between loosely-coupled multimodal and contextual modelling.

More recently, Graph Neural Networks (GNNs) have shown great promise and yielded remarkable improvements in ERC, by revealing rich expressive power of mining structural information and data relations [17, 18]. A routine solution is to construct a heterogeneous graph where each modality of an utterance is regarded as a node, and connected with other modalities of the same utterance as well as connected with the utterances in same modality in the same dialogue. Carefully-tweaked edge-weighting strategies usually follow. On this basis, multimodal and contextual dependencies among utterances can be modelled simultaneously through message passing, and thus deliver tighter entanglement and richer interaction. Powerful as these GNN-based methods are, they still suffer from two limitations:

i) **Insufficient multivariate relationships**. Conventional GNNs assume pairwise relationships of objects of interest, and can only offer an approximation of higher-order and multivariate relationships through multiple pairs [1, 10]. However, degeneration of those multivariate relationships into pairwise formulation may harm the expressiveness [20,30]. Therefore, complex multivariate relationships in ERC may not be sufficiently modelled by previous GNN-based methods.

ii) **Underestimated high-frequency information**. It has been shown that the propagation rule of GNNs (i.e., aggregating and smoothing messages from neighbours) is an analogy to a fixed low-pass filter [26, 31], and it is mainly low-frequency messages that flow in the graph while the effects of high-frequency ones are much weakened. Moreover, Bo *et al*. [2] show that low-frequency messages, which retain the commonality of node features, perform better on assortative graphs (in which the linked nodes tend to have similar features and share the same label). In contrast, high-frequency information that mirrors discrepancy and inconsistency is more crucial on disassortative graphs. For ERC, the constructed graphs are in general highly disassortative, where inconsistent emotional messages may exist among modalities (say being sarcastic) or short-term context. Hence, high-frequency information may provide crucial guidance, which is however badly ignored by previous GNN-based ERC models, incurring bottleneck of performance improvement.

To address these issues, in this work we propose **M**ultivariate **M**ulti-frequency **M**ultimodal Graph Neural **Net**work (M^{3}Net), which aims to capture more sufficient multivariate relationships among modalities and context, while benefiting from multi-frequency information within the graph. At the core of M^{3}Net are two parallel components, multivariate propagation and multi-frequency propagation. Concretely, we first construct a hypergraph neural network with edge-dependent node weights [7] for multivariate propagation, in which each modality of an utterance is represented as a node. We construct multimodal and contextual hyperedges, which can connect arbitrary number of nodes, and thus can naturally encode relationships of higher arity. Meanwhile, we model multi-frequency information upon an undirected GNN, by adapting a set of frequency filters [2, 8] to distil different frequency constituents from the node features. We adaptively integrate different frequency signals to capture the varying importance of emotion discrepancy and emotion commonality in the local neighbourhood, so as to achieve adaptive information sharing pattern.

The effectiveness of our work is further demonstrated by extensive experimental studies on two popular multimodal ERC datasets IEMOCAP [3] and MELD [27]. We show that M^{3}Net outperforms previous state-of-the-art methods.

2. Related work

2.1. Emotion recognition in conversation

Due to the great potential in interactive applications, Emotion Recognition in Conversation (ERC) has attracted great interests of many researchers. Various attempts have been made to study multimodal and contextual relationships in ERC. Some early works [13,14,24] focused more on contextual dependencies and conducted simple feature concatenation to perform multimodal modelling. To enhance the interrelation between modalities and context, recent methods introduced more advanced schemes such as positional attention [34] and adaptive computation [5]. However, these methods still encode multimodal and contextual relationships in a loosely-coupled manner, which may result in weak interaction between them. More recently, some researchers formulated the ERC task upon GNNs, which are powerful in mining data relations hence exhibit superior capability to model contextual and multimodal dependencies [17, 18]. Nevertheless, these GNN-based models still deliver insufficient multivariate relationships and underrate high-frequency signals, as we discussed.

In this work, we propose a new approach that enhances multivariate information among modalities and context, whilst capturing the varying importance of emotion discrepancy and emotion commonality, to deliver more sufficient multimodal and contextual modelling.

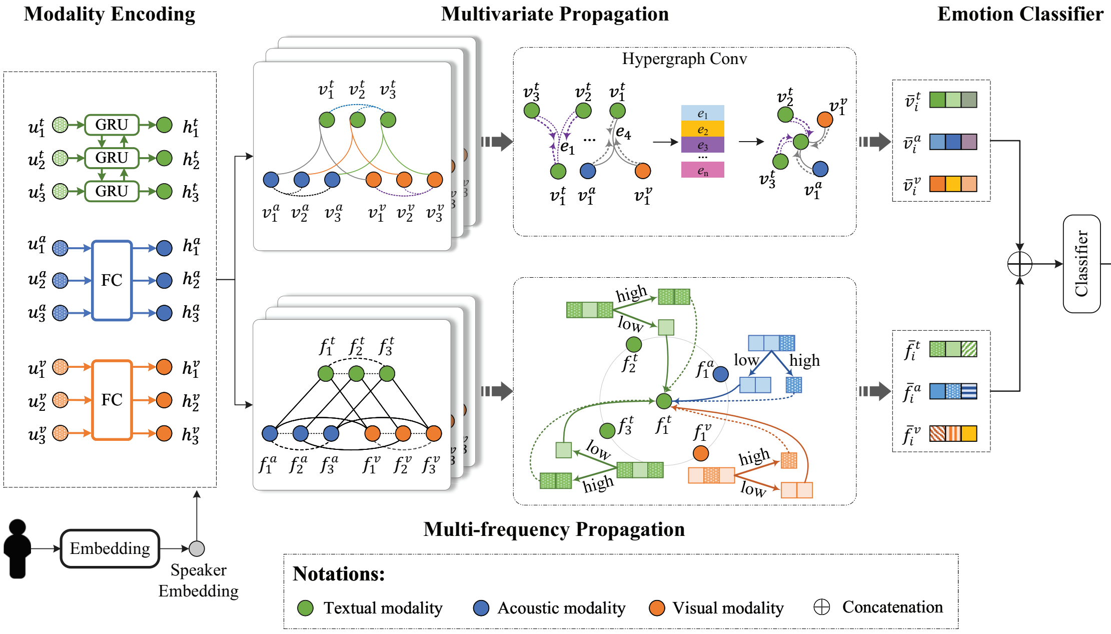


Figure 2. Detailed architecture of the proposed M^{3}Net.

2.2. Graph neural networks

Graph Neural Networks (GNNs) have a distinct advantage in modelling data relationships, and have been widely employed in various applications such as recommendation [16] and action recognition [6]. GNNs have also inspired ERC researchers and offer a new solution for the ERC task, from unimodal setting [12, 28] to multimodal scenario [17, 18]. However, previous works fail to address the general limits of GNNs, including conducting pairwise formulation and erasing high-frequency information, which motivates our work. We present a GNN-based model that encodes relationships of higher arity and values different frequency signals in the neighbourhood. We empower GNNs to better capture the inherent relations among utterances and boost up the performance.

3. Methodology

In a nutshell, an ERC model aims to detect the emotion state of each utterance in a dialogue. Formally, a dialogue contains a sequence of N utterances $\{(u_1, p_1), (u_2, p_2), ..., (u_N, p_N)\}$, where each utterance u_i, spoken by speaker p_i, consists of multi-sensory data, including textual (u_i^t), visual (u_i^v) and acoustic (u_i^a) modalities. The goal is to predict the emotion category of each constituent utterance u_i from a predefined set of C classes.

Figure 2 shows the architecture of the proposed M^{3}Net. In general, M^{3}Net contains four components: modality encoding, multivariate propagation, multi-frequency propagation, and an emotion classifier.

3.1. Modality encoding

A conversation is sequential in nature and consists of multiple speakers. Therefore, we firstly process unimodal utterances with speaker and context information, to obtain speaker- and context-aware unimodal representations. Specifically, we denote each speaker with a one-hot vector s_i and maintain a lookup table for M speakers to calculate the speaker embedding S_i at the i-th conversation turn:

$$S_i = W_s s_i, \tag{1}$$

in which $S_i \in \mathbb{R}^{D_h}$ and W_s is trainable weight. In addition, we employ a bidirectional Gated Recurrent Unit (GRU) to encode the conversational textual features. We empirically observe that encoding visual and acoustic modalities with recurrent modules has no positive effect on the performance, hence use two one-hidden-layer multilayer perceptrons W_1 and W_2 to encode acoustic and visual modalities respectively. Mathematically,

$$\begin{aligned} c_i^t &= \overleftrightarrow{GRU}(u_i^t, c_{i(+,-)1}^t), \\ c_i^a &= W_1 u_i^a + b_i^a, \\ c_i^v &= W_2 u_i^v + b_i^v, \end{aligned} \tag{2}$$

in which $c_i^t, c_i^a, c_i^v \in \mathbb{R}^{D_h}$. We then add speaker embedding to obtain speaker- and context-aware unimodal representations $\{h_i^t, h_i^a, h_i^v\}$ at the i-th conversation turn:

$$h_i^x = c_i^x + S_i, \quad x \in \{t, a, v\}. \tag{3}$$

3.2. Multivariate propagation

The main idea of the multivariate propagation module is to explore the multivariate and high-order information among multiple modalities and conversational context. We begin by constructing a hypergraph $\mathcal{H}$ with edge-dependent node weights, from the sequentially encoded utterances.

3.2.1 Graph construction

Generally, given a sequence of utterances with N conversation turns, we construct a hypergraph $\mathcal{H} = (\mathcal{V}_\mathcal{H}, \mathcal{E}_\mathcal{H}, \omega, \gamma)$, in which each node $v \in \mathcal{V}_\mathcal{H}$ ($|\mathcal{V}_\mathcal{H}| = 3N$) corresponds to a unimodal utterance, and every hyperedge $e \in \mathcal{E}_\mathcal{H}$ ($|\mathcal{E}_\mathcal{H}| = 3 + N$) encodes multimodal or contextual dependencies. A weight $\omega(e)$ is assigned for every hyperedge $e \in \mathcal{E}_\mathcal{H}$, and a weight $\gamma_e(v)$ for every hyperedge $e \in \mathcal{E}_\mathcal{H}$ and every node v incident to e. Let $\mathbf{H} \in \mathbb{R}^{|\mathcal{V}_\mathcal{H}| \times |\mathcal{E}_\mathcal{H}|}$ represent the incidence matrix, in which a nonzero entry $H_{ve} = 1$ indicates that the hyperedge e is incident with the node v; otherwise $H_{ve} = 0$.

Nodes: Each modality of an utterance is represented as a node in hypergraph, i.e., v_i^t for the textual modality, v_i^a for the acoustic modality and v_i^v for the visual modality. We initialize the node embeddings $\{v_i^t, v_i^a, v_i^v\}$ with the sequentially encoded representations $\{h_i^t, h_i^a, h_i^v\}$ respectively.

Hyperedges: The design of hyperedges is based on the assumption that the emotion behind an utterance in a dialogue is mainly determined by the joint effect of multiple modalities and conversational context, and multivariate relationships may exist across both dimensions. To fully investigate the complex multivariate relationships, we construct multimodal hyperedges and contextual hyperedges for each node. Concretely, as shown in Figure 2, each node v_i^x ($x \in \{t, a, v\}$) is firstly connected to all other utterances in the same modality in the same dialogue $\{v_j^x | j \in [1, N], j \neq i\}$, with one contextual hyperedge. Moreover, each node v_i^x is connected to other modalities of the same utterances $\{v_i^z | z \in \{t, a, v\}, z \neq x\}$, with one multimodal hyperedge. In this fashion, the constructed hypergraph is able to capture high-order and multivariate messages that are beyond pairwise formulation.

Weights: Unlike previous GNN-based ERC models [12, 18] which manually tweak the edge weighting strategies with complicated relation learning or similarity metrics, we use randomly initialized weight values to avoid complicating our model. Specifically, we define two types of weights in the hypergraph: i) an edge weight $\omega(e)$ for every hyperedge e, and ii) a node weight $\gamma_e(v)$ for every hyperedge e incident to v, a.k.a., edge-dependent node weight [7]. Intuitively, $\gamma_e(v)$ measures the contribution of node v to hyperedge e, and thus reinforces fine-grained multimodal and contextual dependencies. Edge-dependent node weights can thus be represented by a weighted incidence matrix $\hat{\mathbf{H}} \in \mathbb{R}^{|\mathcal{V}_\mathcal{H}| \times |\mathcal{E}_\mathcal{H}|}$:

$$\hat{\mathbf{H}} = \begin{cases} \gamma_e(v), \text{ if hyperedge } e \text{ is incident with node } v; \\ \quad 0, \text{ otherwise.} \end{cases} \tag{4}$$

3.2.2 Neighbour aggregation

We reformulate hypergraph convolution operation [1] to propagate multivariate embeddings. We also remove feature transformation at each iteration as it is observed to be of little benefit. Specifically, we first conduct node convolution by aggregating node features to update hyperedge embeddings, and then conduct hyperedge convolution to spread the hyperedge messages to the nodes. Mathematically,

$$\mathbf{V}^{(l+1)} = \sigma(\mathbf{D}_\mathcal{H}^{-1} \mathbf{H} \mathbf{W}_e \mathbf{B}^{-1} \hat{\mathbf{H}}^\top \mathbf{V}^{(l)}), \tag{5}$$

in which $\mathbf{V}^{(l)} = \{v_{i,(l)}^x | i \in [1, N], x \in \{t, a, v\}\} \in \mathbb{R}^{|\mathcal{V}_\mathcal{H}| \times D_h}$ is the input at layer l. σ is a non-linear activation function. $\mathbf{W}_e = \text{diag}(w(e_1), ..., w(e_{|\mathcal{E}_\mathcal{H}|}))$ is the hyperedge weight matrix. $\mathbf{D}_\mathcal{H} \in \mathbb{R}^{|\mathcal{V}_\mathcal{H}| \times |\mathcal{V}_\mathcal{H}|}$ and $\mathbf{B} \in \mathbb{R}^{|\mathcal{E}_\mathcal{H}| \times |\mathcal{E}_\mathcal{H}|}$ are the node degree matrix and hyperedge degree matrix, respectively. By this means, the high-order multimodal and contextual relationships are gradually refined. After L iterations, we get the outputs of the last iteration $v_{i,(L)}^x$ as the multivariate representations:

$$\overline{v_i^t} = v_{i,(L)}^t, \ \overline{v_i^a} = v_{i,(L)}^a, \ \overline{v_i^v} = v_{i,(L)}^v. \tag{6}$$

3.3. Multi-frequency propagation

The above multivariate propagation module is able to capture high-order dependencies that are beyond pairwise, but it still follows the generic graph learning protocol which aggregates and smooths signals from the local neighbourhood. This can be interpreted as a form of low-pass filter and the smoothness of messages is basically spreading low-frequency information whilst erasing high-frequency information [2, 26, 31]. However, as discussed earlier, high-frequency information that mirrors emotion discrepancy of nodes may be crucial for ERC, and combining the power of messages with varying frequencies is worth exploring. It thus motivates us to propose a multi-frequency propagation module to distil different frequency constituents with varying importance. For this purpose, we further construct an undirected graph $\mathcal{G} = (\mathcal{V}_\mathcal{G}, \mathcal{E}_\mathcal{G})$ from the sequentially encoded utterances, in parallel with the multivariate module.

3.3.1 Graph construction

We construct an undirected graph $\mathcal{G} = (\mathcal{V}_\mathcal{G}, \mathcal{E}_\mathcal{G})$ whose nodes $\mathcal{V}_\mathcal{G}$ are identical to the ones in $\mathcal{H}$, denoted with $\{f_i^t, f_i^a, f_i^v\}$. The node embeddings at the first iteration are initialized with the sequentially encoded representations $\{h_i^t, h_i^a, h_i^v\}$ as well. Different from $\mathcal{H}$, we construct a set of edges $\mathcal{E}_\mathcal{G}$ with pairwise connections. Similarly, we connect each node f_i^x to all other utterances in the same modality in the same dialogue $\{f_j^x | j \in [1, N], j \neq i\}$, as well as to other modalities of the same utterances $\{f_i^z | z \in \{t, a, v\}, z \neq x\}$. The constructed graph $\mathcal{G}$ is shown in Figure 2, with adjacency matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}_\mathcal{G}| \times |\mathcal{V}_\mathcal{G}|}$. The normalized graph Laplacian matrix can be represented as $\mathbf{L} = \mathbf{I} - \mathbf{D}_\mathcal{G}^{-1/2} \mathbf{A} \mathbf{D}_\mathcal{G}^{-1/2}$, where $\mathbf{D}_\mathcal{G} \in \mathbb{R}^{|\mathcal{V}_\mathcal{G}| \times |\mathcal{V}_\mathcal{G}|}$ is a diagonal degree matrix and $\mathbf{I}$ is an identity matrix.

3.3.2 Multi-frequency filtering

We first design a low-pass filter $\mathcal{F}_l$ and a high-pass filter $\mathcal{F}_h$ to distil the signals from the node features:

$$\begin{aligned} \mathcal{F}_l &= \mathbf{I} + \mathbf{D}_\mathcal{G}^{-1/2} \mathbf{A} \mathbf{D}_\mathcal{G}^{-1/2} = 2\mathbf{I} - \mathbf{L}, \\ \mathcal{F}_h &= \mathbf{I} - \mathbf{D}_\mathcal{G}^{-1/2} \mathbf{A} \mathbf{D}_\mathcal{G}^{-1/2} = \mathbf{L}. \end{aligned} \tag{7}$$

It can be noticed that the high-pass filter is equivalent to the normalized graph Laplacian matrix, which is consistent with the theory in image signal processing that the Laplacian kernel can be employed to highlight high-frequency edge information. According to theory of graph Fourier transform [2, 29], given a signal φ, the filtering operation by $\mathcal{F}_l$ and $\mathcal{F}_h$ can be regarded as the convolutional $*_C$ between φ and corresponding convolutional kernels:

$$\mathcal{F}_l *_C \varphi = \mathcal{F}_l \cdot \varphi, \ \mathcal{F}_h *_C \varphi = \mathcal{F}_h \cdot \varphi. \tag{8}$$

3.3.3 Graph learning

After obtaining low-pass and high-pass filters, we leverage the filters to adaptively aggregate messages with varying frequencies. Specifically, we use a weighted sum to combine low-frequency and high-frequency messages:

$$\begin{aligned} \mathbf{F}^{(k+1)} &= \mathbf{R}^l(\mathcal{F}_l \cdot \mathbf{F}^{(k)}) + \mathbf{R}^h(\mathcal{F}_h \cdot \mathbf{F}^{(k)}) \\ &= \mathbf{F}^{(k)} + (\mathbf{R}^l - \mathbf{R}^h) \mathbf{D}_\mathcal{G}^{-1/2} \mathbf{A} \mathbf{D}_\mathcal{G}^{-1/2} \mathbf{F}^{(k)}, \end{aligned} \tag{9}$$

in which $\mathbf{F}^{(k)} = \{f_{i,(k)}^x | i \in [1, N], x \in \{t, a, v\}\} \in \mathbb{R}^{|\mathcal{V}_\mathcal{G}| \times D_h}$ is the input at layer k. $\mathbf{R}^l, \mathbf{R}^h \in \mathbb{R}^{|\mathcal{V}_\mathcal{G}| \times |\mathcal{V}_\mathcal{G}|}$ are the weight matrices for low-frequency and high-frequency information, respectively. Eq. (9) can be written in another form as

$$f_{i,(k+1)} = f_{i,(k)} + \sum_{j \in \mathcal{N}_i} \frac{r_{ij}^l - r_{ij}^h}{\sqrt{|\mathcal{N}_j|}\sqrt{|\mathcal{N}_i|}} f_{j,(k)}, \tag{10}$$

where $\mathcal{N}_i$ is the neighbouring nodes of node i. r_{ij}^l and r_{ij}^h are the weight contributions of node j's low-frequency and high-frequency signals to node i, respectively, and they meet the constraint $r_{ij}^l + r_{ij}^h = 1$.

To effectively learn the coefficient $r_{ij}^l - r_{ij}^h$ in Eq. (10), we follow FAGCN [2] to employ a self-gating mechanism, which considers the correlation between the central node and neighbours:

$$r_{ij}^l - r_{ij}^h = \tanh(W_3(f_{i,(k)} \oplus f_{j,(k)})). \tag{11}$$

Here, $\oplus$ is the concatenation operation and $W_3 \in \mathbb{R}^{2D_h \times 1}$ is a trainable weight matrix. $\tanh(\cdot)$ is the hyperbolic tangent function that scales the value in $[-1, 1]$. By this means, the coefficient $r_{ij}^l - r_{ij}^h$ can readily model the varying importance of different frequency constituents. For instance, if $r_{ij}^l - r_{ij}^h < 0$, the high-frequency messages dominate, and node i receives the discrepancy between node i and the neighbour j (i.e., $f_{i,(k)} - f_{j,(k)}$); and it holds vice versa.

Now we gradually spread the multi-frequency information over the graph. By stacking K layers, each node receives the multi-frequency signals from K-hop neighbours, and we use outputs of the final layer as the multi-frequency representations:

$$\overline{f_i^t} = f_{i,(K)}^t, \ \overline{f_i^a} = f_{i,(K)}^a, \ \overline{f_i^v} = f_{i,(K)}^v. \tag{12}$$

3.3.4 Differences with FAGCN

The graph learning rule of the above multi-frequency module is closely related to Frequency Adaptation Graph Convolutional Networks (FAGCN) [2], which proposes to adaptively integrate low-frequency and high-frequency signals as well. Although we derive inspiration from FAGCN, our multi-frequency module contains several critical distinctions: (i) FAGCN introduces a hyper-parameter to balance the identity matrix and Laplacian matrix when defining filters while our method is hyper-parameter free; (ii) FAGCN always updates node embeddings based on the inputs at first layer, while we gradually refine the node embeddings based on the outputs of previous layer. We present the performance comparison between our multi-frequency module and FAGCN in Section 5.5 and show by extensive experiments that our design outperforms FAGCN.

3.4. Emotion classification

The emotion classifier takes as input the concatenated multivariate and multi-frequency representations to perform emotion prediction. Mathematically,

$$e_i = \overline{v_i^t} \oplus \overline{f_i^t} \oplus \overline{v_i^a} \oplus \overline{f_i^a} \oplus \overline{v_i^v} \oplus \overline{f_i^v}, \tag{13}$$

where e_i is the emotion representation for utterance i, and contains both multivariate dependencies and multi-frequency information. Finally, we feed e_i into a softmax layer to obtain the emotion category:

$$\begin{aligned} \tilde{e}_i &= \mathrm{ReLU}(e_i), \\ \mathcal{P}_i &= \mathrm{softmax}(W_4 \tilde{e}_i + b_4), \\ \hat{y}_i &= \underset{\tau}{\mathrm{argmax}}(\mathcal{P}_i[\tau]), \end{aligned} \tag{14}$$

where W_4 is trainable weight, $\mathcal{P}_i \in \mathbb{R}^C$ and $\hat{y}_i$ is the predicted label for utterance u_i.

3.5. Training objective

We follow prior works [18, 24] to use categorical cross-entropy along with L_2-regularization as the loss function:

$$L = -\frac{1}{\sum_{s=1}^{Num} c(s)} \sum_{i=1}^{Num} \sum_{j=1}^{c(i)} log\mathcal{P}_{i,j}[y_{i,j}] + \lambda \left\|\theta\right\|_2 , \quad (15)$$

where Num is the number of dialogues, $c(i)$ is the number of utterances in dialogue i, $\mathcal{P}_{i,j}$ and $y_{i,j}$ are the probabilistic distribution of class labels and the ground-truth label for utterance j in dialogue i, respectively. λ is the L_2-regularizer weight and θ denotes the trainable parameters in the model.

4. Experiments

4.1. Datasets

We compare the performance of our proposed M^{3}Net against prior works on two popular multimodal datasets, IEMOCAP [3] and MELD [27], following dominant data split protocol and modality employment as in previous works [5, 17, 18].

IEMOCAP contains 151 dyadic dialogues of ten speakers and 7,433 utterances labelled with one of six emotion categories: happy, sad, neutral, angry, excited, or frustrated. We use 120 dialogues with 5,810 utterances for training and validation, and the rest for testing. We employ language, video and audio modalities for emotion prediction.

MELD is a multiparty emotional conversational dataset which is collected from the TV show *Friends*. MELD contains 1,433 rounds of conversations and 13,708 utterances. Each utterance is annotated as one of seven emotion labels: anger, disgust, sadness, joy, surprise, fear, or neutral. We use 1,039 dialogues with 9,989 utterances for training, 114 dialogues with 1,109 utterances for validation, and the rest for testing. We follow previous works [17, 18] to employ language, video and audio modalities.

4.2. Unimodal feature extraction

In this paper, we use pre-extracted unimodal features following identical settings in previous studies [5, 11, 24].

The textual features are extracted using the RoBERTa Large model [22], which is firstly fine-tuned for emotion prediction from the transcript of conversations. After the fine-tuning process, the utterances are fed to the model and the activations from the final four layers are extracted as four textual vectors, which are then normalized and averaged for the final textual representation. The dimension of textual features in our paper is 1024.

The acoustic features are obtained by the openSMILE toolkit [9]. The visual features are extracted with a pre-trained DenseNet [19] for the MELD dataset, and through a 3D-CNN for the IEMOCAP dataset. More details are stated in appendix.

Dataset	Batch	Optimizer	D_h	L	K	Dropout
IEMOCAP	16	Adam (lr=1e-4)	512	3	4	0.5
MELD	16	Adam (lr=1e-4)	512	3	3	0.4

Table 1. Details of hyper-parameters in our experiments.

4.3. Baselines

For a comprehensive evaluation of M^{3}Net, we compare our model with the following state-of-the-art methods:

- **CMN** [14] seeks to model contextual information from dialogue history. It uses two GRUs for two speakers and stores contexts as memories. It is not applicable to multiparty scenarios, hence no results on MELD.
- **ICON** [13] is an extension of CMN, which connects outputs from speaker GRUs in CMN with another GRU, so as to explicitly model inter-speaker interaction. Similar to CMN, ICON is not applicable to multiparty scenarios, hence no results on MELD.
- **DialogueRNN** [24] employs three GRU cells to respectively keep track of global context, speaker state and emotion state throughout the conversation. It is capable of handling multiparty conversations.
- **MetaDrop** [5] introduces a binary maintain-or-drop decision learning mechanism to learn adaptive fusion paths, as well as simultaneously capture multimodal and contextual relations.
- **DialogueGCN** [12] uses graph relational modelling to encode context. Each utterance is represented as a node, and connected with other nodes in the same dialogue within a context window. It originally focuses on textual modality and we extend it to multimodal scenario by concatenating the unimodal embeddings.
- **MMGCN** [18] constructs a heterogeneous graph by regarding each modality of each utterance as a node. It designs separate edge weighting mechanisms for inter-modal and intra-modal edges, and encodes both multimodal and contextual information with deep layers.
- **MM-DFN** [17] proposes a graph-based dynamic fusion module to keep track of conversational context in different semantic spaces, and enhance complementarity between modalities.

4.4. Settings and evaluation metrics

The proposed model is implemented using PyTorch and torch-geometric packages. The networks are trained on a machine with 1 NVIDIA GeForce RTX 3090. We follow dominant evaluation protocols to use accuracy and F1-score as the metrics to measure the performance. Paired t-test is performed to test the significance of performance improvement with a default significance level of 0.05. Models are

	Methods	Network	IEMOCAP Average (w)		MELD Average (w)	
			Accuracy	F1	Accuracy	F1
GloVe	CMN$^{\triangleleft}$ [14]	Non-GNN	-	58.50	-	-
	ICON* [13]	Non-GNN	64.00	63.50	-	-
	DialogueRNN† [24]	Non-GNN	63.51	62.90	59.92	57.60
	MetaDrop$^{\diamondsuit}$ [5]	Non-GNN	65.76	65.58	-	58.30
	DialogueGCN† [12]	GNN-based	66.17	66.24	57.01	55.59
	MMGCN† [18]	GNN-based	65.80	65.41	60.42	58.31
	MM-DFN† [17]	GNN-based	68.21	68.18	59.81	58.42
	M^{3}Net (ours)	GNN-based	**69.50**	**69.08**	**61.65**	**59.22**
RoBERTa	DialogueGCN† [12]	GNN-based	63.96	64.44	63.49	62.78
	MMGCN† [18]	GNN-based	66.79	66.99	66.63	65.13
	DialogueRNN$^{\diamondsuit}$ [24]	Non-GNN	68.64	68.72	65.94	65.31
	MetaDrop$^{\diamondsuit}$ [5]	Non-GNN	69.38	69.59	66.63	66.30
	MM-DFN† [17]	GNN-based	69.87	69.48	67.01	66.17
	M^{3}Net (ours)	GNN-based	**72.46**	**72.49**	**68.28**	**67.05**

Table 2. Comparison with previous state-of-the-art methods on IEMOCAP and MELD. Bold font denotes the best performances. Average(w) = weighted average. $^{\triangleleft}$ from [24]; * from [13]; $^{\diamondsuit}$ from [5]; † from our reimplementation using open source codes.

trained using Adam [21] with a batch size of 16 on both datasets. We test L and K in the range from 1 to 7 and present the best-performing results. Full details of hyperparameters for both datasets are shown in Table 1. Code is available at https://github.com/feiyuchen7/M3NET.

5. Results and analysis

5.1. Comparison with state-of-the-arts

We contrast our model with a wide range of state-of-the-art methods in Table 2. It can be seen that on both datasets, our proposed M^{3}Net surpasses previous methods and achieves new state-of-the-art records in terms of both metrics of accuracy and F1-score. In particular, M^{3}Net outperforms previous GNN-based methods, including DialogueGCN, MMGCN and MM-DFN, which manually tweak edge weighting strategies with complicated relation learning or similarity metrics, to capture multimodal and contextual relationships. We suggest that the advantage of our method is due to the investigation into multivariate and multi-frequency information among modalities and context, which is neglected by previous methods.

5.2. Textual features from BERT vs. GloVe

As stated in Section 4.2, in this work the inputting textual features are extracted from a pre-trained RoBERTa Large model, which according to our observation, can boost up the performance compared with traditional GloVe-based textual features. In order to verify whether our model can deliver good performance regardless of the sources of textual features, we further conduct experiments using GloVe embeddings and present comparison with previous methods. The results are shown in Table 2. It can be observed that M^{3}Net outperforms other baselines based on either textual feature source, through which we can infer that our multivariate and multi-frequency modelling delivers major improvements.

	Methods	IEMOCAP		MELD	
		Acc.	F1	Acc.	F1
	M^{3}Net	72.46	72.49	68.28	67.05
1	w/o multivariate info.	70.06	70.05	67.74	66.36
2	w/o multi-frequency info.	69.87	69.74	67.36	66.03
3	w/o hyperedge weight $\omega(e)$	70.30	70.45	68.11	66.99
4	w/o node weight $\gamma_e(v)$	70.98	71.02	68.05	66.92
5	w/o both weights	70.12	70.09	67.89	66.75
6	$\mathcal{H} \rightarrow \mathcal{G}$ in series	68.39	68.44	68.20	66.84
7	$\mathcal{G} \rightarrow \mathcal{H}$ in series	69.50	69.70	68.05	66.85

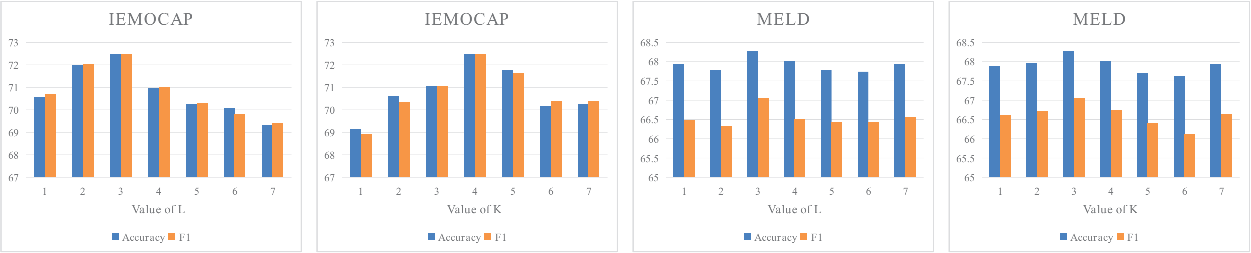
Table 3. Ablation studies of M^{3}Net.

5.3. Ablation studies

To gain better insights to the constituents of our model, we perform ablation studies on the key components of M^{3}Net and present results in Table 3.

Effect of multivariate information. We first explore the effect of multivariate information among modalities and context. To achieve this, we remove the multivariate propagation module (i.e., the hypergraph $\mathcal{H}$) and perform classification based on multi-frequency representations only, shown as variant 1 in Table 3. Under this setting, we can observe a decrease of 2.40% in accuracy and 2.44% in F1-score on IEMOCAP, as well as a decrease of 0.54% in accuracy and 0.69% in F1-score on MELD. This demonstrates the effectiveness of introducing multivariate propagation, which can naturally encode relationships of higher arity.

Effect of multi-frequency information. Another core component of M^{3}Net is the multi-frequency propagation module. Similarly, we test the importance of this module by



(a) Effect of L on IEMOCAP (b) Effect of K on IEMOCAP (c) Effect of L on MELD (d) Effect of K on MELD

Figure 3. Results of M^{3}Net at different graph layers. In (a) and (c), effects of L are tested by fixing K as in the best-performing models. In (b) and (d), effects of K are tested by fixing L as in the best-performing models.

having it removed and performing predictions using multivariate representations only. Variant 2 reports the results of this configuration, from which a sharp degradation of performance can be observed. This stands as a convincing proof of the validity of introducing different frequency information into ERC, which can guide the model to capture the varying importance of emotion discrepancy and emotion commonality in the local neighbourhood.

Effect of weights in the hypergraph. In Section 3.2.1, we define two types of weights in hypergraph $\mathcal{H}$ to capture the multivariate relationships in a fine-grained level. We hence conduct experiments to verify the effect of these two weights. It can be seen from variants 3 to 5 that removing either or both weights (i.e., setting weight value $\omega(e)$ or/and $\gamma_e(v)$ as 1) harms the performance on both datasets. This indicates that the formulated weights benefit the training.

Effect of parallel modelling. In M^{3}Net, we propagate multivariate and multi-frequency information in parallel. We further conduct experiments to compare it with two-step serial modelling and show the results as variants 6 and 7. Serial modelling slightly reduces the performance on MELD but leads to dramatic decreases on IEMOCAP, which implies the effectiveness of the parallel modelling.

5.4. Discussions on graph layers

M^{3}Net contains two parallel graphs, and the graph propagation plays a pivotal role. To investigate the impact of stacking different graph layers, we conduct a grid search on the number of layers. Specifically, we search the layer numbers of multivariate propagation (L) and multi-frequency propagation (K) in the range from 1 to 7 and summarize the results in Figure 3. On IEMOCAP, the effects of L and K are similar. At first, the results steadily improve as stacking more layers, and peak at $L = 3$ and $K = 4$ respectively. Further stacking more layers has little positive impact on the performance. On the other hand, it can be noticed that the results on MELD are less sensitive to the number of graph layers, with no special pattern, as shallow or deep layers can all yield decent performance.

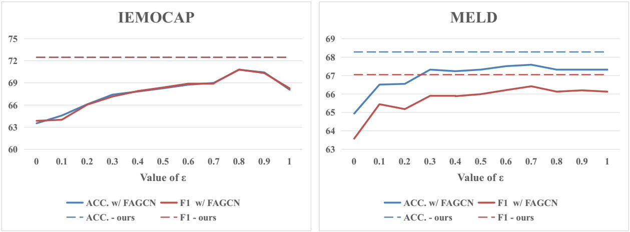


Figure 4. Performance comparison with FAGCN.

5.5. Comparison with FAGCN

As stated in Section 3.3.4, the graph propagation rule of our multi-frequency module is closely related to FAGCN [2] but retains critical distinctions. To further demonstrate the effectiveness of our method, we present an additional comparison with FAGCN. Specifically, we maintain the multivariate module, and replace our multi-frequency modelling strategy (Eq. (7) to Eq. (11)) with the one introduced in FAGCN. Since FAGCN introduces a hyper-parameter $\epsilon \in [0, 1]$ when defining filters, we test ϵ in the range of $[0, 1]$ with a step of 0.1. The comparison is summarized in Figure 4. Apparently, ϵ is a vital factor and dramatically impacts the performance, especially on IEMOCAP. However, under no circumstances can these variants with FAGCN outperform the original M^{3}Net. This indicates the superiority of our multi-frequency modelling mechanism.

6. Conclusion

This paper proposes a GNN-based model to address the ERC problem. We present Multivariate Multi-frequency Multimodal Graph Neural Network (M^{3}Net) to investigate the multivariate relationships among modalities and context, and take full advantage of different frequency information which reflects emotion discrepancy and commonality. Extensive experimental results on two datasets show the superiority of our model.

References

[1] Song Bai, Feihu Zhang, and Philip H. S. Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognit.*, 110:107637, 2021. 2, 4

[2] Deyu Bo, Xiao Wang, Chuan Shi, and Huawei Shen. Beyond low-frequency information in graph convolutional networks. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 3950–3957, 2021. 2, 4, 5, 8

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359, 2008. 2, 6

[4] Feiyu Chen, Jie Shao, Anjie Zhu, Deqiang Ouyang, Xueliang Liu, and Heng Tao Shen. Modeling hierarchical uncertainty for multimodal emotion recognition in conversation. *IEEE Trans. Cybern.*, 2022. 2

[5] Feiyu Chen, Zhengxiao Sun, Deqiang Ouyang, Xueliang Liu, and Jie Shao. Learning what and when to drop: Adaptive multimodal and contextual dynamics for emotion recognition in conversation. In *MM '21: ACM Multimedia Conference*, pages 1064–1073, 2021. 1, 2, 6, 7

[6] Hyung-Gun Chi, Myoung Hoon Ha, Seung-geun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 20154–20164, 2022. 3

[7] Uthsav Chitra and Benjamin J. Raphael. Random walks on hypergraphs with edge-dependent vertex weights. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pages 1172–1181, 2019. 2, 4

[8] Yushun Dong, Kaize Ding, Brian Jalaian, Shuiwang Ji, and Jundong Li. Adagnn: Graph neural networks with adaptive frequency response filter. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management*, pages 392–401, 2021. 2

[9] Florian Eyben, Martin Wöllmer, and Björn W. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th International Conference on Multimedia 2010*, pages 1459–1462, 2010. 6

[10] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 3558–3565, 2019. 2

[11] Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. COSMIC: commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, 2020. 6

[12] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 154–164, 2019. 1, 3, 4, 6, 7

[13] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. ICON: interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, 2018. 1, 2, 6, 7

[14] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, Volume 1 (Long Papers)*, pages 2122–2132, 2018. 2, 6, 7

[15] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In *MM '20: The 28th ACM International Conference on Multimedia*, pages 1122–1131, 2020. 1

[16] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020*, pages 639–648, 2020. 3

[17] Dou Hu, Xiaolong Hou, Lingwei Wei, Lian-Xin Jiang, and Yang Mo. MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022*, pages 7037–7041, 2022. 2, 3, 6, 7

[18] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers)*, pages 5666–5675, 2021. 1, 2, 3, 4, 6, 7

[19] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 2261–2269, 2017. 6

[20] Yuchi Huang, Qingshan Liu, and Dimitris N. Metaxas. Video object segmentation by hypergraph cut. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 1738–1745, 2009. 2

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015. 7

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 6

[23] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Trans. Affect. Comput.*, 2022. 1

[24] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. Dialoguernn: An attentive RNN for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 6818–6825, 2019. 1, 2, 6, 7

[25] Yuzhao Mao, Qi Sun, Guang Liu, Xiaojie Wang, Weiguo Gao, Xuan Li, and Jianping Shen. Dialoguetrm: Exploring the intra- and inter-modal emotional behaviors in the conversation. *CoRR*, abs/2010.07637, 2020. 2

[26] Hoang NT and Takanori Maehara. Revisiting graph neural networks: All we have is low-pass filters. *CoRR*, abs/1905.09550, 2019. 2, 4

[27] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, pages 527–536, 2019. 2, 6

[28] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers)*, pages 1551–1560, 2021. 1, 3

[29] David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag.*, 30(3):83–98, 2013. 5

[30] Xiangguo Sun, Hongzhi Yin, Bo Liu, Hongxu Chen, Jiuxin Cao, Yingxia Shao, and Nguyen Quoc Viet Hung. Heterogeneous hypergraph embedding for graph classification. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining*, pages 725–733, 2021. 2

[31] Felix Wu, Amauri H. Souza Jr., Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pages 6861–6871, 2019. 2, 4

[32] Liwen Xu, Zhengtao Wang, Bin Wu, and Simon Lui. MDAN: multi-level dependent attention network for visual emotion analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 9469–9478, 2022. 1

[33] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multitask learning for multimodal sentiment analysis. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 10790–10797, 2021. 1

[34] Dong Zhang, Weisheng Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Modeling both intra- and inter-modal influence for real-time emotion detection in conversations. In *MM '20: The 28th ACM International Conference on Multimedia*, pages 503–511, 2020. 2