MM-DFN: MULTIMODAL DYNAMIC FUSION NETWORK FOR EMOTION RECOGNITION IN CONVERSATIONS

Dou Hu[1] *Xiaolong Hou*[1] *Lingwei Wei*[2] *Lianxin Jiang*[1] *Yang Mo*[1]

[1] Ping An Life Insurance Company of China, Ltd.
[2] Institute of Information Engineering, Chinese Academy of Sciences

ABSTRACT

Emotion Recognition in Conversations (ERC) has considerable prospects for developing empathetic machines. For multimodal ERC, it is vital to understand context and fuse modality information in conversations. Recent graph-based fusion methods generally aggregate multimodal information by exploring unimodal and cross-modal interactions in a graph. However, they accumulate redundant information at each layer, limiting the context understanding between modalities. In this paper, we propose a novel Multimodal Dynamic Fusion Network (MM-DFN) to recognize emotions by fully understanding multimodal conversational context. Specifically, we design a new graph-based dynamic fusion module to fuse multimodal context features in a conversation. The module reduces redundancy and enhances complementarity between modalities by capturing the dynamics of contextual information in different semantic spaces. Extensive experiments on two public benchmark datasets demonstrate the effectiveness and superiority of the proposed model.

***Index Terms*—** emotion recognition, emotion recognition in conversations, multimodal fusion, dialogue systems

1. INTRODUCTION

Emotion Recognition in Conversations (ERC) aims to detect emotions in each utterance of the conversation. It has considerable prospects for developing empathetic machines [1]. This paper studies ERC under a multimodal setting, *i.e.*, acoustic, visual, and textual modalities.

A conversation often contains rich contextual clues [2, 3], which are essential for identifying emotions. The key success factors of multimodal ERC are accurate context understanding and multimodal fusion. Previous context-dependent works [3–5] model conversations as sequence or graph structures to explore contextual clues within a single modality. Although these methods can be naturally extended multimodal paradigms by performing early/late fusion such as [6–8], it is difficult to capture contextual interactions between modalities, which limits the utilization of multiple modalities. Besides, some carefully-designed hybrid fusion methods [9–12] focus on the alignment and interaction between modalities in isolated or sequential utterances. These methods ignore complex interactions between utterances, resulting in leveraging context information in conversations insufficiently.

Recent remarkable works [13, 14] model unimodal and cross-modal interactions in a graph structure, which provides complementarity between modalities for tracking emotions. However, these graph-based fusion methods aggregate contextual information in a specific semantic space at each layer, gradually accumulating redundant information. It limits context understanding between modalities. The contextual information continuously aggregated can be regarded as specific views where each view can have its individual representation space and dynamics. We believe that modeling these dynamics of contextual information in different semantic spaces can reduce redundancy and enhance complementarity, accordingly boosting context understanding between modalities.

In this paper, we propose a novel Multimodal Dynamic Fusion Network (MM-DFN) to recognize utterance-level emotion by sufficiently understanding multimodal conversational context. Firstly, we utilize a modality encoder to track speaker states and context in each modality. Secondly, inspired by [15, 16], we improve the graph convolutional layer [17] with gating mechanisms and design a new Graph-based Dynamic Fusion (GDF) module to fuse multimodal context information. The module utilizes graph convolution operation to aggregation context information of both inter- and intra-modality in a specific semantic space at each layer. Meanwhile, the gating mechanism is used to learn the intrinsic sequential patterns of contextual information in adjacent semantic space. The GDF module can control information flow between layers, reducing redundancy and promoting the complementarity between modalities. The stack of GDFs can naturally fuse multimodal context features by embedding them into a dynamic semantic space. Finally, an emotion classifier is used to predict the emotion label of the utterance.

We conduct a series of experiments on two public benchmark datasets, *i.e., IEMOCAP* and *MELD*. Results consistently demonstrate that MM-DFN significantly outperforms comparison methods. The main contributions are summarized as follows: 1) We propose a novel MM-DFN to facilitate multimodal context understanding for ERC. 2) We design a new

Corresponding author. Email: HUDOU470@pingan.com.cn

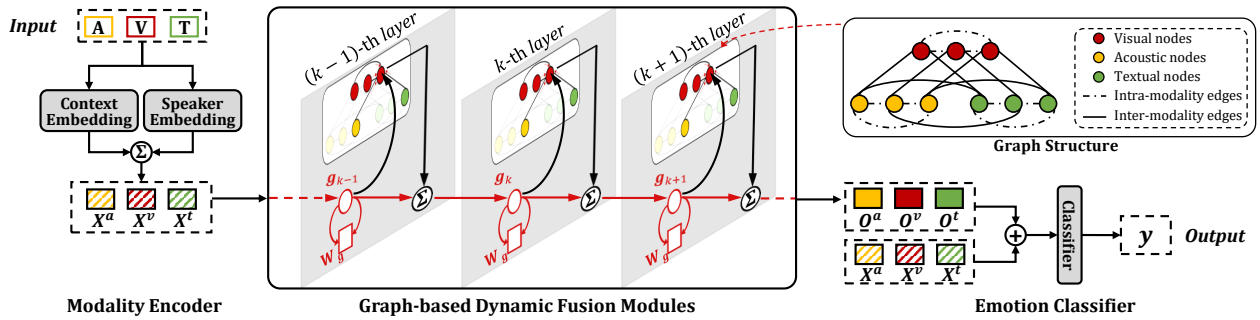

Fig. 1. The architecture of the proposed MM-DFN. Given input multimodal features, modality encoder first captures features of context and speaker in each modality. Then, in each conversation, we construct the fully connected graph in each modality, and connect nodes corresponding to the same utterance between different modalities. Based on the graph, graph-based dynamic fusion modules are stacked to fuse multimodal context features, dynamically and sequentially. Finally, based on the concatenation of features, an emotion classifier is applied to identify emotion label of each utterance.

graph-based dynamic fusion module to fuse multimodal conversational context. This module can reduce redundancy and enhance complementarity between modalities. 3) Extensive experiments on two benchmark datasets demonstrate the effectiveness and superiority of the proposed model[1].

2. METHODOLOGY

Formally, given a conversation $U = [u_1, ..., u_N]$, $u_i = \{\mathbf{u}_i^a, \mathbf{u}_i^v, \mathbf{u}_i^t\}$, where N is the number of utterances. $\mathbf{u}_i^a, \mathbf{u}_i^v, \mathbf{u}_i^t$ denote the raw feature representation of u_i from the acoustic, visual, and textual modality, respectively. There are M speakers $P = \{p_1, ..., p_M\}(M \geq 2)$. Each utterance u_i is spoken by the speaker $p_{\phi(u_i)}$, where ϕ maps the index of the utterance into the corresponding speaker. Moreover, we define U_λ to represent the set of utterances spoken by the party p_λ. $U_\lambda = \{u_i | u_i \in U$ and u_i spoken by $p_\lambda, \forall i \in [1, N]\}, \lambda \in [1, M]$. The goal of multimodal ERC is to predict the emotion label y_i for each utterance u_i from pre-defined emotions $\mathcal{Y}$.

In this section, we propose a novel Multimodal Dynamic Fusion Network (MM-DFN) to fully understand the multimodal conversational context for ERC, as shown in Fig. 1.

2.1. Modality Encoder

To capture context features for the textual modality, we apply a bi-directional gated recurrent unit (BiGRU); for the acoustic and visual modalities, we apply a fully connected network. The context embedding can be computed as:

$$\begin{aligned} \mathbf{c}_i^\varsigma &= \mathbf{W}_c^\varsigma \mathbf{u}_i^\varsigma + \mathbf{b}_c^\varsigma, \varsigma \in \{a, v\}, \\ \mathbf{c}_i^t, \mathbf{h}_i^c &= \overleftrightarrow{GRU}_c(\mathbf{u}_i^t, \mathbf{h}_{i-1}^c), \end{aligned} \tag{1}$$

where $\overleftrightarrow{GRU}_c$ is a BiGRU to obtain context embeddings and $\mathbf{h}_i^c$ is the hidden vector. $\mathbf{W}_c^a, \mathbf{W}_c^v, \mathbf{b}_c^a, \mathbf{b}_c^v$ are trainable parameters. Considering the impact of speakers in a conversation, we also employ a shared-parameter BiGRU to encode different contextual information from multiple speakers:

$$\mathbf{s}_i^\delta, \mathbf{h}_{\lambda,j}^s = \overleftrightarrow{GRU}_s(\mathbf{u}_i^\delta, \mathbf{h}_{\lambda,j-1}^s), j \in [1, |U_\lambda|], \delta \in \{a, v, t\}, \tag{2}$$

where $\overleftrightarrow{GRU}_s$ indicates a BiGRU to obtain speaker embeddings. $\mathbf{h}_{\lambda,j}^s$ is the j-th hidden state of the party p_λ. $\lambda = \phi(u_i)$. U_λ refers to all utterances of p_λ in a conversation.

2.2. Graph-based Dynamic Fusion Modules

2.2.1. Graph Construction

Following [13], we build an undirected graph to represent a conversation, denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ refers to a set of nodes. Each utterance can be represented by three nodes for differentiating acoustic, visual, and textual modalities. Given N utterances, there are $3N$ nodes in the graph. We add both context embedding and speaker embedding to initialize the embedding of nodes in the graph:

$$\mathbf{x}_i^\delta = \mathbf{c}_i^\delta + \gamma^\delta \mathbf{s}_i^\delta, \delta \in \{a, v, t\}, \tag{3}$$

where $\gamma^a, \gamma^v, \gamma^t$ are trade-off hyper-parameters. $\mathcal{E}$ refers to a set of edges, which are built based on two rules. The first rule is that any two nodes of the same modality in the same conversation are connected. The second rule is that each node is connected with nodes corresponding to the same utterance but from different modalities. Following [18], edge weights are computed as: $\mathbf{A}_{ij} = 1 - \frac{\arccos(\text{sim}(\mathbf{x}_i, \mathbf{x}_j))}{\pi}$, where $\text{sim}(\cdot)$ is cosine similarity function.

2.2.2. Dynamic Fusion Module

Based on the graph, we improve [17] with gating mechanisms to fuse multimodal context features in the conversation. We utilize graph convolution operation to aggregate context information of both inter- and intra-modality in a specific semantic space at each layer. Meanwhile, inspired by [15], we

[1]The code is available at https://github.com/zerohd4869/MM-DFN

leverage gating mechanisms to learn intrinsic sequential patterns of contextual information in different semantic spaces. The updating process using gating mechanisms is defined as:

$$\begin{aligned}
&\mathbf{\Gamma}_{\varepsilon}^{(k)} = \sigma(\mathbf{W}_{\varepsilon}^{g} \cdot [\mathbf{g}^{(k-1)}, \mathbf{H}'^{(k-1)}] + \mathbf{b}_{\varepsilon}^{g}), \varepsilon = \{u, f, o\}, \\
&\tilde{\mathbf{C}}^{(k)} = \tanh(\mathbf{W}_{C}^{g} \cdot [\mathbf{g}^{(k-1)}, \mathbf{H}'^{(k-1)}] + \mathbf{b}_{C}^{g}), \\
&\mathbf{C}^{(k)} = \mathbf{\Gamma}_{f}^{(k)} \odot \mathbf{C}^{(k-1)} + \mathbf{\Gamma}_{u}^{(k)} \odot \tilde{\mathbf{C}}^{(k)}, \ \mathbf{g}^{(k)} = \mathbf{\Gamma}_{o}^{(k)} \odot \tanh(\mathbf{C}^{(k)}),
\end{aligned} \tag{4}$$

where $\mathbf{\Gamma}_{u}^{(k)}, \mathbf{\Gamma}_{f}^{(k)}, \mathbf{\Gamma}_{o}^{(k)}$ refer to the update gate, the forget gate, and the output gate in the k-th layer, respectively. $\mathbf{g}^{(0)}$ is initialized with zero. $\mathbf{W}_{\Gamma}^{g}, \mathbf{b}_{\Gamma}^{g}$ are learnable parameters. $\sigma(\cdot)$ is a sigmoid function. $\tilde{\mathbf{C}}^{(k)}$ stores contextual information of previous layers. The update gate $\mathbf{\Gamma}_{u}^{(k)}$ controls what part of the contextual information is written to the memory, while the forget gate $\mathbf{\Gamma}_{f}^{(k)}$ decides what redundant information in $\mathbf{C}^{(k)}$ is deleted. The output gate $\mathbf{\Gamma}_{o}^{(k)}$ reads selectively for passing into a graph convolution operation. Following [16], the modified convolution operation can be defined as:

$$\mathbf{H}^{(k)} = \text{ReLU}\left(((1-\alpha)\tilde{\mathbf{P}}\mathbf{H}'^{(k-1)} + \alpha\mathbf{H}^{(0)})((1-\beta_{k-1})\mathbf{I}_n + \beta_{k-1}\mathbf{W}^{(k-1)})\right), \tag{5}$$

where $\tilde{\mathbf{P}} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$ is the graph convolution matrix with the renormalization trick. α, β_k are two hyperparameters. $\beta_k = log(\frac{\rho}{k} + 1)$. ρ is also a hyperparameter. $\mathbf{W}^{(k)}$ is the weight matrix. $\mathbf{H}^{(0)}$ is initialized with $\mathbf{X}^a, \mathbf{X}^v, \mathbf{X}^t$. $\mathbf{I}_n$ is an identity mapping matrix. Then, the output of k-th layer can be computed as, $\mathbf{H}'^{(k)} = \mathbf{H}^{(k)} + \mathbf{g}^{(k)}$.

2.3. Emotion Classifier

After the stack of K layers, representations of three modalities for each utterance i can be refined as $\mathbf{o}_i^a, \mathbf{o}_i^v, \mathbf{o}_i^t$. Finally, a classifier is used to predict the emotion of each utterance:

$$\hat{\mathbf{y}}_i = \text{Softmax}(\mathbf{W}_z[\mathbf{x}_i^a; \mathbf{x}_i^v; \mathbf{x}_i^t; \mathbf{o}_i^a; \mathbf{o}_i^v; \mathbf{o}_i^t] + \mathbf{b}_z), \tag{6}$$

where $\mathbf{W}_z$ and $\mathbf{b}_z$ are trainable parameters. We apply cross-entropy loss along with L2-regularization to train the model:

$$\mathcal{L} = -\frac{1}{\sum_{l=1}^{L}\tau(l)}\sum_{i=1}^{L}\sum_{j=1}^{\tau(i)}\mathbf{y}_{i,j}^{l} log(\hat{\mathbf{y}}_{i,j}^{l}) + \eta\|\Theta\|_2, \tag{7}$$

where L is the total number of samples in the training set. $\tau(i)$ is the number of utterances in sample i. $\mathbf{y}_{i,j}^{l}$ and $\hat{\mathbf{y}}_{i,j}^{l}$ denote the one-hot vector and probability vector for emotion class j of utterance i of sample l, respectively. Θ refers to all trainable parameters. η is the L2-regularization weight.

3. EXPERIMENTS

3.1. Datasets

IEMOCAP [19] contains dyadic conversation videos between pairs of ten unique speakers. It includes 7,433 utterances and 151 dialogues. Each utterance is annotated with one of six emotion labels. We follow the previous studies [5, 13] that use the first four sessions for training, use the last session for testing, and randomly extract 10% of the training dialogues as validation split. ***MELD*** [2] contains multi-party conversation videos collected from Friends TV series, where two or more speakers are involved in a conversation. It contains 1,433 conversations, 13,708 utterances and 304 different speakers. Each utterance is annotated with one of seven emotion labels. For a fair comparison, we conduct experiments using the predefined train/validation/test splits in MELD.

3.2. Comparison Methods

TFN [9] and **LMF** [10] make non-temporal multimodal fusion by tensor product. **MFN** [11] synchronizes multimodal sequences using a multi-view gated memory. **bc-LSTM** [6] leverages an utterance-level LSTM to capture multimodal features. **ICON** [7], an extension of CMN [20], provides conversational features from modalities by multi-hop memories. **DialogueRNN** [4] introduces a recurrent network to track speaker states and context during the conversation. **DialogueCRN** [3] designs multi-turn reasoning modules to understand conversational context. **DialogueGCN** [5] utilizes graph structures to combine contextual dependencies. **MMGCN** [13] uses a graph-based fusion module to capture intra- and inter- modality contextual features. All baselines are reproduced under the same environment, except [7], which is only applicable for dyadic conversation and the results are from the original paper. Because [3–5] are designed for unimodal ERC, a early concatenation fusion is introduced to capture multimodal features in their implementations.

Implementation Details. Following [13], raw utterance-level features of acoustic, visual, and textual modality are extracted by TextCNN [21], *OpenSmile* [22], and DenseNet [23], respectively. We use focal loss [24] for training due to the class imbalance. The number of layers K are 16 and 32 for IEMOCAP and MELD. α is set to 0.2 and ρ is set to 0.5.

3.3. Experimental Results and Analysis

Overall Results and Ablation Study. The overall results are reported in Table 1. MM-DFN consistently obtains the best performance over the comparison methods on both datasets, which shows the superiority of our model. Table 2 shows ablation studies by removing key components of the proposed model. When removing either the graph-based dynamic fusion (GDF) module or speaker embedding (Speaker), the results decline significantly on both datasets. When further removing the context embedding (Context), the results decrease further. It shows the effectiveness of the three components.

Comparison with Different Fusion Modules. After the modality encoder, we replace GDF with the following six fusion modules: **Concat/Gate Fusion**, **Tensor/Memory Fusion** [10, 11], **Early/Late Fusion + GCN** [13], and **Graph-based Fusion** (GF) [13]. From Table 3, GF and GDF outperform all fusion modules in the first block since the two

Methods	IEMOCAP								MELD						
	Happy	*Sad*	*Neutral*	*Angry*	*Excited*	*Frustrated*	Acc	w-F1	*Neutral*	*Surprise*	*Sadness*	*Happy*	*Anger*	Acc	w-F1
TFN [9]	37.26	65.21	51.03	54.64	58.75	56.98	55.02	55.13	77.43	47.89	18.06	51.28	44.15	60.77	57.74
LMF [10]	37.76	66.53	52.39	57.53	58.41	59.27	56.50	56.49	76.97	47.06	21.15	54.20	46.64	61.15	58.30
MFN [11]	48.19	73.41	56.28	63.04	64.11	61.82	61.24	61.60	77.27	48.29	23.24	52.63	41.32	60.80	57.80
bc-LSTM [6]	33.82	78.76	56.75	64.35	60.25	60.75	60.51	60.42	75.66	48.57	22.06	52.10	44.39	59.62	57.29
ICON [7]	32.80	74.40	60.60	68.20	68.40	66.20	64.00	63.50	-	-	-	-	-	-	-
DialogueRNN [4]	32.20	80.26	57.89	62.82	73.87	59.76	63.52	62.89	76.97	47.69	20.41	50.92	45.52	60.31	57.66
DialogueCRN [3]	**53.23**	**83.37**	62.96	66.09	75.40	66.07	67.16	67.21	77.01	50.10	26.63	52.77	45.15	61.11	58.67
DialogueGCN [5]	51.57	80.48	57.69	53.95	72.81	57.33	63.22	62.89	75.97	46.05	19.60	51.20	40.83	58.62	56.36
MMGCN [13]	45.14	77.16	64.36	68.82	74.71	61.40	66.36	66.26	76.33	48.15	**26.74**	53.02	46.09	60.42	58.31
MM-DFN	42.22	78.98	**66.42***	**69.77***	**75.56***	**66.33***	**68.21***	**68.18***	**77.76***	**50.69***	22.93	**54.78**	**47.82***	**62.49***	**59.46***

Table 1. Results under the multimodal setting (A+V+T). We present the overall performance of Acc and w-F1, which mean the overall accuracy score and weighted-average F1 score, respectively. We also report F1 score per class, except two classes (i.e. *Fear* and *Disgust*) on MELD, whose results are not statistically significant due to the smaller number of training samples. Best results are highlighted in bold. * represents statistical significance over state-of-the-art scores under the paired-t test ($p < 0.05$).

Methods	IEMOCAP	MELD
MM-DFN	68.18	59.46
- w/o GDF - w Speaker - w Context	63.80	58.50
- w GDF - w/o Speaker - w Context	66.89	58.45
- w/o GDF - w/o Speaker - w Context	62.90	58.50
- w/o GDF - w/o Speaker - w/o Context	54.81	58.08

Table 2. Ablation results of MM-DFN. We report w-F1 score for both datasets.

Fusion Modules	IEMOCAP	MELD
Concat / Gate Fusion	63.80 / 64.30	58.50 / 57.87
Tensor / Memory Fusion	61.05 / 65.51	58.54 / 58.48
Early / Late Fusion + GCN	64.19 / 65.34	58.69 / 58.43
Graph-based Fusion (GF)	67.02	58.54
- w/o Inter-Modal - w Intra-Modal	66.91	58.53
- w Inter-Modal - w/o Intra-Modal	66.11	58.29
Graph-based Dynamic Fusion (GDF)	68.18	59.46
- w/o Inter-Modal - w Intra-Modal	67.82	59.15
- w Inter-Modal - w/o Intra-Modal	66.22	58.31

Table 3. Results against different fusion modules. We report w-F1 score for both datasets.

graph-based fusion modules sufficiently capture intra- and inter-modality interactions in conversations, which provides complementarity between modalities. GDF achieves better performance, reducing redundancy and promoting the complementarity between modalities, which shows the superiority of multimodal fusion. Besides, for GF and GDF, we analyze the impact of inter- and intra-modality edges in the graph for fusion. Intra-/Inter-Modal refers to building edges according to the first/second rule. Ignoring any rules can hurt performance in GF and GDF, which shows that modeling contextual interactions of both inter- and intra-modality, can better utilize the complementarity between modalities. Compared with GF, GDF obtains a better performance in all variants. It shows that GDF can reduce both inter- and intra-modality redundancies and fuse multimodal context better.

Comparison under Different Modality Settings. Table 4 shows the results of MM-DFN and the GF-based variant under different modality settings. As expected, bimodal and trimodal models outperform the corresponding unimodal models on both datasets. Under unimodal types, textual modality performs better than acoustic and visual. Under bimodal types, GDF outperforms GF consistently. It again confirms the superiority of GDF. Meanwhile, under acoustic and textual modalities (A+T), both GF and GDF achieve the best performance over other bimodal types, which indicates a stronger complementarity between rich textual semantics and affective audio features. GDF can reduce redundancy as well as enhance complementarity between modalities and thus obtain better results. Moreover, under acoustic and visual modalities (A+V), GDF outperforms GF by a large margin. This phenomenon reflects that the acoustic and visual features have high entanglement and redundancy, limiting the performance of GF. Our GDF encourages disentangling and reduces redundancy by controlling information flow between modalities, accordingly obtaining better fusion representations.

Modality	IEMOCAP		MELD	
	GF	**GDF**	GF	**GDF**
A / V / T	-	47.79 / 27.46 / 61.07	-	42.72 / 32.34 / 56.95
A + V	54.73	56.35	42.74	44.67
A + T	65.03	65.41	57.85	58.34
V + T	62.07	62.63	57.78	58.49
A + V + T	67.02	68.18	58.54	59.46

Table 4. Results of graph-based fusion methods under different modality settings. Fusion modules are not used under unimodal types. We report w-F1 score for both datasets.

4. CONCLUSION

This paper proposes a Multimodal Dynamic Fusion Network (MM-DFN) to fully understand conversational context for multimodal ERC task. A graph-based dynamic fusion (GDF) module is designed to fuse multimodal features in a conversation. The stack of GDFs learns dynamics of contextual information in different semantic spaces, successfully reducing redundancy and enhancing complementarity between modalities. Extensive experiments on two benchmark datasets demonstrate the effectiveness and superiority of MM-DFN.

5. REFERENCES

[1] Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, et al., "A survey on empathetic dialogue systems," *Inf. Fusion*, vol. 64, pp. 50–70, 2020.

[2] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, et al., "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *ACL*, 2019, pp. 527–536.

[3] Dou Hu, Lingwei Wei, and Xiaoyong Huai, "Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations," in *ACL/IJCNLP*, 2021, pp. 7042–7052.

[4] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, et al., "Dialoguernn: An attentive RNN for emotion detection in conversations," in *AAAI*, 2019, pp. 6818–6825.

[5] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, et al., "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," in *EMNLP/IJCNLP*, 2019, pp. 154–164.

[6] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, et al., "Context-dependent sentiment analysis in user-generated videos," in *ACL*, 2017, pp. 873–883.

[7] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, et al., "ICON: interactive conversational memory network for multimodal emotion detection," in *EMNLP*, 2018, pp. 2594–2604.

[8] Yahui Fu, Shogo Okada, Longbiao Wang, et al., "CONSK-GCN: conversational semantic- and knowledge-oriented graph convolutional network for multimodal emotion recognition," in *ICME*. 2021, pp. 1–6, IEEE.

[9] Amir Zadeh, Minghai Chen, Soujanya Poria, et al., "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017, pp. 1103–1114.

[10] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, et al., "Efficient low-rank multimodal fusion with modality-specific factors," in *ACL (1)*. 2018, pp. 2247–2256, ACL.

[11] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, et al., "Memory fusion network for multi-view sequential learning," in *AAAI*, 2018, pp. 5634–5641.

[12] Feiyu Chen, Zhengxiao Sun, Deqiang Ouyang, et al., "Learning what and when to drop: Adaptive multimodal and contextual dynamics for emotion recognition in conversation," in *ACM Multimedia*, 2021, pp. 1064–1073.

[13] Jingwen Hu, Yuchen Liu, Jinming Zhao, et al., "MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation," in *ACL/IJCNLP*, 2021, pp. 5666–5675.

[14] Jiaxing Liu, Sen Chen, Longbiao Wang, et al., "Multimodal emotion recognition with capsule graph convolutional based representation fusion," in *ICASSP*, 2021, pp. 6339–6343.

[15] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] Ming Chen, Zhewei Wei, Zengfeng Huang, et al., "Simple and deep graph convolutional networks," in *ICML*, 2020, pp. 1725–1735.

[17] Thomas N. Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR (Poster)*, 2017.

[18] Konstantinos Skianis, Fragkiskos D. Malliaros, and Michalis Vazirgiannis, "Fusing document, collection and label graph-based representations with word embeddings for text classification," in *TextGraphs@NAACL-HLT*, 2018, pp. 49–58.

[19] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, et al., "IEMOCAP: interactive emotional dyadic motion capture database," *Lang. Resour. Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[20] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, et al., "Conversational memory network for emotion recognition in dyadic dialogue videos," in *NAACL-HLT*, 2018, pp. 2122–2132.

[21] Yoon Kim, "Convolutional neural networks for sentence classification," in *EMNLP*, 2014, pp. 1746–1751.

[22] Björn W. Schuller, Anton Batliner, Stefan Steidl, et al., "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, no. 9-10, pp. 1062–1087, 2011.

[23] Gao Huang, Zhuang Liu, Laurens van der Maaten, et al., "Densely connected convolutional networks," in *CVPR*, 2017, pp. 2261–2269.

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, et al., "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988.