Interspeech 2024
1-5 September 2024, Kos, Greece

MM-NodeFormer: Node Transformer Multimodal Fusion for Emotion Recognition in Conversation

Zilong Huang, Man-Wai Mak, Kong Aik Lee

Dept. of Electrical and Electronic Engineering,
The Hong Kong Polytechnic University, Hong Kong SAR, China

zi-long.huang@connect.polyu.hk, man.wai.mak@polyu.edu.hk, kong-aik.lee@polyu.edu.hk

Abstract

Emotion Recognition in Conversation (ERC) has great prospects in human-computer interaction and medical consultation. Existing ERC approaches mainly focus on information in the text and speech modalities and often concatenate multimodal features without considering the richness of emotional information in individual modalities. We propose a multimodal network called MM-NodeFormer for ERC to address this issue. The network leverages the characteristics of different Transformer encoding stages to fuse the emotional features from the text, audio, and visual modalities according to their emotional richness. The module considers text as the main modality and audio and visual as auxiliary modalities, leveraging the complementarity between the main and auxiliary modalities. We conducted extensive experiments on two public benchmark datasets, IEMOCAP and MELD, achieving an accuracy of 74.24% and 67.86%, respectively, significantly higher than many state-of-the-art approaches.

Index Terms: emotion recognition in conversation, multimodal network, feature fusion

1. Introduction

A fundamental difference between humans and machines is their ability in expressing and interpreting emotion. The primary objective of emotion recognition in conversation (ERC) is to discern and assign emotion labels to individual utterances within a conversation. Essentially, ERC is a classification task, aiming to assign an emotion category to each expression in a dialogue from a predefined set of emotion categories [1].

A primary characteristic of ERC is the interdependencies among the utterances in a conversation, which could be contextual- [2] or speaker-dependent [3]. Prior studies have put forward diverse session-based approaches to capturing the dependency. These approaches use bi-directional LSTM [4], multi-layer GRU [5], and graph-based models [6] on multimodal inputs. Although they can capture the contextual dependency, they often simply concatenate the features from multiple modalities without finding a better representation through advanced fusion methods. This direct concatenation of features leads to high-dimensional feature space, causing sub-optimal performance [7]. Although there have been advancements in feature fusion [4, 6], most approaches treat the features of different modalities identically without considering the emotion content in different modalities. Also, these feature-level fusion methods do not consider the interaction between different modalities, limiting the recognition performance of the fused features.

This work was supported by the RGC of Hong Kong SAR, Grant No. PolyU 15210122

To effectively utilize the emotion contents in multiple modalities and to consider their interaction, we propose a novel multimodal fusion network called MM-NodeFormer, which leverages the characteristics of different Transformer encoding stages to perform feature fusion. Specifically, the audio and visual features extracted from Wav2vec2.0 [8] and CLIP [9] in a dialog are aligned with the self-attended RoBERTa [10] text features. Then, multiple Transformer encoders capture the dependence among these dialog-level feature sequences. Subsequently, an emotion classifier is employed to predict the emotion labels.

The contributions of our works are summarized as follows: 1) we propose a novel MM-NodeFormer for ERC, featuring a multi-stage fusion strategy to improve the quality of audio-visual-text multimodal fusion; 2) we design a new Transformer-based multimodal fusion called NodeFormer, allocating weights based on the richness of emotion information to extract deeper emotional features; and 3) extensive experiments on IEMOCAP and MELD demonstrate the effectiveness and superiority of the proposed model.

2. Related Work

Unlike traditional emotion recognition, ERC uses contextual relationships between the utterances in a conversation. Therefore, it is crucial to capture such relationship via contextual modeling. To this end, the conversational memory network [5] and its extension, ICON [2], use two GRUs (one for each speaker) to learn the inter-speaker influences and dynamics in dyadic conversational videos. DialogRNN [11] assumes that an utterance's emotional state depends on its speaker, the context in the preceding utterances of all speakers, and the emotional states of preceding utterances. It uses three GRUs to model these relationships. DialogGCN [6] acquires contextual dependencies by employing a graph neural network. MMGCN [12] utilizes a multimodal graph-based fusion module to capture the contextual features. MM-DFN [13] aggregates contextual information within and between modalities in specific semantic spaces using graph convolution operations. M2FNet [14] incorporates a multi-head fusion attention layer to seamlessly merge features extracted from diverse modalities. UniMSE [15] focuses on the similarities and complementarities between emotion at syntactic and semantic levels. EmotionIC [16] models a conversation at the feature extraction and classification levels. CFN-ESA [17] models emotional transfer by introducing an emotional shift module and extracts the shifting information through an auxiliary task. EmoCaps [18] introduces the concept of emotion vectors into multi-modal emotion recognition.

The models above either only use single-modal information to perform ERC or do not assess the emotional content of fea-

10.21437/Interspeech.2024-538

tures from different modalities to leverage the dominant modality. To fill this gap, we propose using a multi-modality Transformer fusion model with learnable fusion weights to capture the contextual relationship in the utterances of a conversation.

3. Problem Statement

Assume that a dialogue has k utterances $\mathcal{U} = \{u_i\}_{i=1}^{k}$ following a temporal sequence and that the utterances have emotion labels $\{y_i\}_{i=1}^{k}$ and speaker labels $\{s_i\}_{i=1}^{k}$. Here, $y_i \in \mathcal{Y}$, and $\mathcal{Y}$ denotes the set of emotion labels. Each utterance u_i is associated with a video clip, an audio segment, and a text transcript. Mathematically, a dialogue can be represented as:

$$\{\mathcal{U}, \mathcal{Y}\} = \Big\{ \{u_i^{(\delta)}, y_i\} \text{ s.t. } \delta \in \{a, t, v\} \text{ and } i \in \{1, \ldots, k\} \Big\}. \tag{1}$$

To simplify the notations, we denote $u_i^{(\delta)}$ in (1) as the δ-type input of the i^{th} utterance, which can be in the form of text (t), audio (a), and video (v), respectively. We assume that there are S participants $\mathcal{P} = \{P_1, \ldots, P_S; S \geq 2\}$ in a dialogue. The i^{th} utterance u_i is spoken by participant P_s such that $s = \phi(u_i) \in \{1, \ldots, S\}$, where $\phi()$ is a mapping function. ERC aims to output an emotion label $\hat{y}_i$ associated with the i^{th} utterance u_i in the utterance sequence $\mathcal{U} = (u_1, u_2, \ldots, u_k)$ with S participants, using the information contained in the k utterances.

4. Methodology

4.1. Overall Architecture

Figure 1 illustrates the dataflow and structure of the proposed MM-NodeFormer, which consists of three processing stages.

- *Utterance- and Context-Levels Multimodal Feature Extraction:* We extract utterance-level features independently from the text, audio, and visual modalities using their respective pre-trained models. We then determine the dependencies of the features using bi-directional GRUs.
- *Dialogue-Level Feature Fusion:* We propose a fusion module called NodeFormer to combine the audio, text, and visual modalities.
- *Emotion Classification:* The final representations after fusion are fed into a classifier to predict the emotion labels in the dialogue.

4.2. Utterance-Level Features Extraction

Each utterance in a dialog is associated with multimodal inputs u_i^t, u_i^a , and u_i^v for text, audio, and visual, respectively, where $i \in \{1, \ldots, k\}$. The input from each modality is processed separately using a pre-trained feature extractor specific to that modality. Specifically, we use RoBERTa [10], Wav2vec2.0 [8], and CLIP [9] for the text, audio, and visual modalities, respectively. We perform average pooling at the hidden states of the last layer to obtain three modality-specific features:

$$\begin{aligned} \boldsymbol{x}_i^t &= \text{RoBERTa}(u_i^t) \in \mathbb{R}^{D_t} \\ \boldsymbol{x}_i^a &= \text{Wav2Vec2.0}(u_i^a) \in \mathbb{R}^{D_a} \\ \boldsymbol{x}_i^v &= \text{CLIP}(u_i^v) \in \mathbb{R}^{D_v}, \end{aligned} \tag{2}$$

where D_t, D_a, and D_v are the feature dimensions of the text, audio, and visual modalities, respectively.

4.3. Utterance-Level Representation with Context

The features $\{\boldsymbol{x}_i^\delta; i = 1, \ldots, k\}$ are independently extracted for each utterance u_i. Because utterances in a dialogue are inherently sequential, inspired by [4, 19], we employ Bi-GRUs to capture the contextual features in the dialogue:[1]

$$\boldsymbol{R}^\delta = [\boldsymbol{r}_1^\delta, \ldots, \boldsymbol{r}_k^\delta] = \text{Bi-GRU}(\boldsymbol{x}_1^\delta, \ldots, \boldsymbol{x}_k^\delta) \in \mathbb{R}^{D_r \times k}, \tag{3}$$

where D_r is the GRU's output dimension and $\delta \in \{a, t, v\}$.

4.4. Dialog-Level Feature Extraction

The matrices ($\boldsymbol{R}^t, \boldsymbol{R}^a$, and $\boldsymbol{R}^v$), comprising the text, audio, and visual representations of utterances in a dialogue, are separately inputted to the NodeFormer (middle panel of Figure 1).

Because previous studies [20, 21] on multimodal emotion recognition demonstrated that the visual and audio modalities contain less emotional information than the text modality, we define text as the main modality and visual and audio as the auxiliary modalities. For the text modality, we used an attention mechanism to determine the contextual dependencies between the utterances. First, $\boldsymbol{R}^t$ are linearly transformed into value and query matrices [22]:[2]

$$\begin{aligned} \boldsymbol{V} &= \boldsymbol{W}^V \boldsymbol{R}^t = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k] \in \mathbb{R}^{D_L \times k}, \boldsymbol{v}_i \in \mathbb{R}^{D_L \times 1} \\ \boldsymbol{Q} &= \boldsymbol{W}^Q \boldsymbol{R}^t = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k] \in \mathbb{R}^{D_L \times k}, \boldsymbol{q}_i \in \mathbb{R}^{D_L \times 1}. \end{aligned} \tag{4}$$

The attention weights and attended outputs are computed as:

$$\begin{aligned} \boldsymbol{\alpha}^{qv} &= \text{softmax}(\tanh(\boldsymbol{V}^\top \boldsymbol{Q})) \in \mathbb{R}^{k \times k} \\ \boldsymbol{T}^{att} &= \boldsymbol{V} \boldsymbol{\alpha}^{qv} \in \mathbb{R}^{D_L \times k}. \end{aligned} \tag{5}$$

We concatenate the query $\boldsymbol{Q}$ and attended output $\boldsymbol{T}^{att}$ as the output matrix $\boldsymbol{M}$ of the main modality:

$$\boldsymbol{M} = \text{Concate}(\boldsymbol{Q}, \boldsymbol{T}^{att}) = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_k] \in \mathbb{R}^{D_M \times k}, \tag{6}$$

where $D_M = 2D_L$ is the text modality's feature dimension. For the auxiliary modalities, we concatenate the audio feature matrix $\boldsymbol{R}^a$ with the video feature matrix $\boldsymbol{R}^v$ to obtain a representation $\boldsymbol{A}$:

$$\boldsymbol{A} = \text{Concate}(\boldsymbol{R}^a, \boldsymbol{R}^v) = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k] \in \mathbb{R}^{D_A \times k}, \boldsymbol{a}_i \in \mathbb{R}^{D_A \times 1}, \tag{7}$$

where $D_A = D_a + D_v$ represents the auxiliary feature's dimension. We obtain a matrix $\boldsymbol{C}$ before entering the Transformer encoder:

$$\boldsymbol{C} = \text{Concate}(\boldsymbol{M}, \boldsymbol{W}^A \boldsymbol{A}) = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k] \in \mathbb{R}^{D_C \times k}. \tag{8}$$

Inspired by [22], we pass the fused features $\boldsymbol{C}$ to a series of stacked Transformer encoders to learn contextual representations between utterances. The output from the first Transformer encoder is

$$\boldsymbol{F}_1 = \text{E}_1(\boldsymbol{C}) + \boldsymbol{C} + \boldsymbol{W}^F \boldsymbol{M} \in \mathbb{R}^{D_C \times k}, \tag{9}$$

where $\text{E}_1()$ represents the first Transformer encoder, $\boldsymbol{W}^F \in \mathbb{R}^{D_C \times D_M}$ contains trainable weights. When the number of encoders, denoted as n, is greater than or equal to 2, the final output features can be expressed as:

$$\boldsymbol{F}_n = \text{E}_n(\boldsymbol{F}_{n-1}) + \boldsymbol{F}_{n-1} + \boldsymbol{W}^F \boldsymbol{M} = [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_k] \in \mathbb{R}^{D_C \times k}. \tag{10}$$

In (10), $\boldsymbol{F}_n$ denotes the feature matrix of all utterances in the dialog. To prevent the model from neglecting low-level features, we utilize residual connections [23].

[1] We adopt the terminology in [2] where "context" refers to the information across utterances.

[2] In our case, the key and value matrices are the same.

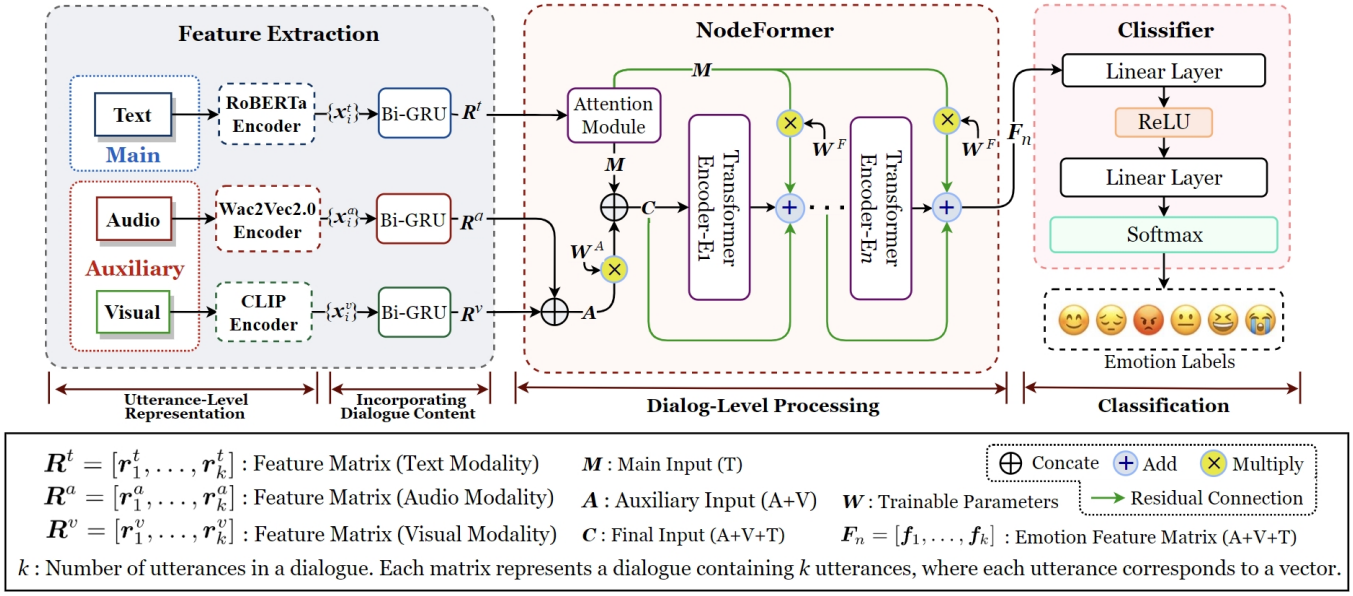

Figure 1: *The three processing stages and architecture of MM-NodeFormer. Solid-line and dashed-line rectangles represent trainable and frozen models, respectively.*

4.5. Emotion Classifier

The final multimodal embeddings $\{\boldsymbol{f}_i\}_{i=1}^{k}$ are passed to a fully connected network:

$$\begin{aligned} \boldsymbol{l}_i &= \mathrm{ReLU}(\boldsymbol{W}_l \boldsymbol{f}_i + \boldsymbol{b}_l) \\ \boldsymbol{p}_i &= \mathrm{softmax}(\boldsymbol{W}_{smax} \boldsymbol{l}_i + \boldsymbol{b}_{smax}), \end{aligned} \tag{11}$$

where $\boldsymbol{p}_i$ contains the probabilities of individual emotion classes of the i^{th} utterance, and $\boldsymbol{W}_l$, $\boldsymbol{W}_{smax}$, $\boldsymbol{b}_l$ and $\boldsymbol{b}_{smax}$ are trainable parameters. We take the emotion label $\hat{y}_i$ with the highest probability as the predicted emotion:

$$\hat{y}_i = \arg\max_j (p_{ij}). \tag{12}$$

5. Experimental Settings

5.1. Datasets

IEMOCAP [24] is a widely used datasets for emotion recognition in conversation. For data partitioning, we adopted the popular "LOSO" (Leave-One-Session-Out) strategy. Since IEMOCAP does not have predefined training and validation splits, we randomly chose 10% of each training split in the LOSO as a validation set. **MELD** [25] is a multi-modal, multi-speaker conversational dataset extracted from the "Friends" TV series. It extends and improves the EmotionLines dataset [26] by extracting the video and audio from the episodes with timestamps aligned with the text in EmotionLines and all utterances in a dialogue belonging to the same scene. To ensure a fair comparison, we followed the predefined training/validation/testing splits provided by the dataset, with data allocation consistent with [13].

5.2. Experimental Setup

The hyperparameter and training settings are shown in Table 1. D_t, D_a and D_v denote the hidden layer dimensions of text, audio and visual features, respectively. D_R is the dimension of Bi-GRU's output, D_C is the dimension of the NodeFormer's output, and n denotes the number of transformer encoder layers within the NodeFormer.

We used categorical cross-entropy loss along with L2 regularization to train the MM-NodeFormer in Figure 1 (excluding the pre-trained modules). We employed an Adam optimizer [27] to train the networks.

6. Results and Analysis

6.1. Overall Results and Effect of Different Modal Settings

Table 2 shows the detailed results of MM-NodeFormer on both datasets. The proposed MM-NodeFormer performs the best among all models, demonstrating its effectiveness.

Table 3 shows the performances using different combinations of modalities. Compared to the bimodal and unimodal settings, the trimodal setting achieves the best performance. Among all the unimodal settings, the text modality achieves the best performance. These results indicate that, for dialogue emotion recognition, the text modality can capture more emotion information compared to the other two modalities. We also conducted an ablation experiment by alternatively setting the text, audio, and visual modalities as the main modalities while setting the other two as the auxiliary modalities. Results in Table 3 show that the NodeFormer fusion module successfully notices the contributions between different modalities, thereby improving the overall quality of the fused features. Table 3 also shows that using a simple Transformer encoder to fuse the concatenated features from the three modalities improves the performance slightly when compared to directly feeding the concatenated features to the classifier. However, the NodeFormer evidently has advantages over the Transformer encoder when we use text as the main modalities.

6.2. Effect of Different Number of Encoders

Table 4 shows the performance of MM-NodeFormer using different numbers of Transformer encoders. The results indicate that when the multimodal features are processed through a stack

Table 1: *The hyperparameter and training settings for IEMOCAP and MELD*

Datasets	D_t	D_a	D_v	D_R	D_C	n	Dropout [28]	L2	Epoch	Batch Size	LR	Criterion
IEMOCAP	768	512	512	200	600	5	0.4	3e-05	200	16	1e-04	Acc&w-F1
MELD	768	512	512	100	300	4	0.4	3e-05	250	70	1e-04	Acc&w-F1

Table 2: *Accuracy and weighted-average F1 score (w-F1) compared with other models in the literature.*

Baseline Model	**Modality Used**	**Proposed Year**	**IEMOCAP**								**MELD**	
			Happy	*Sad*	*Neutral*	*Angry*	*Excited*	*Frustrated*	**Acc**	**w-F1**	**Acc**	**w-F1**
CMN [5]	T	2018	30.38	62.41	52.39	59.83	60.25	60.69	56.56	56.13	-	-
ICON [2]	T	2018	32.80	74.40	60.60	68.20	68.40	66.20	64.00	63.50	-	-
DialogRNN [11]	T	2018	32.20	80.26	57.89	62.82	73.87	59.76	63.52	62.89	-	-
DialogRNN [11]	T+A+V	2018	-	-	-	-	-	-	-	62.90	60.31	57.66
DialogGCN [6]	T	2019	53.23	83.37	62.96	66.09	75.40	66.07	67.16	67.21	58.62	56.36
MMGCN [12]	T+A+V	2021	45.14	77.16	64.36	68.82	74.71	61.40	66.36	66.26	-	58.65
MMDFN [13]	T+A+V	2022	42.22	78.98	66.42	69.77	75.56	66.33	68.21	68.18	62.49	59.46
M2FNet [14]	T+A+V	2022	-	-	-	-	-	-	69.69	69.86	67.85	**66.71**
UniMSE [15]	T+A+V	2022	-	-	-	-	-	-	70.66	70.56	65.09	65.51
EmotionIC [16]	T	2023	-	-	-	-	-	-	69.44	69.61	-	66.32
CFN-ESA [17]	T+A+V	2023	53.67	80.60	**71.65**	70.32	74.82	68.06	71.04	70.78	67.85	66.70
EmoCaps [18]	T+A+V	2022	71.91	**85.06**	64.48	68.99	78.41	66.76	-	71.77	64.00	-
MM-NodeFormer (Proposed)	T+A+V	2023	**76.09**	73.55	69.77	**72.58**	**79.53**	**74.25**	**74.24**	**74.20**	**67.86**	66.09

Table 3: *Performance of using different combinations of modalities and using different modalities as the main (with $*$). "Concat" means concatenating the features ($\boldsymbol{R}$ in Figure 1) from different modalities and feeding the concatenated features directly to the classifier. "TransFormer" means using a simple Transformer encoder to fuse the concatenated features from the three modalities.*

Modality Used	**Fusion Method**	**IEMOCAP**		**MELD**	
		Acc	**w-F1**	**Acc**	**w-F1**
A	—	57.44	58.64	49.87	42.51
V	—	46.33	42.81	48.15	31.30
T	—	71.69	71.73	65.65	63.49
A + V	Concat	57.25	57.45	50.06	43.73
T + V	Concat	72.07	72.16	65.76	64.50
T + A	Concat	72.82	72.82	66.03	64.37
A + T + V	Concat	73.20	73.20	66.22	65.27
A + T + V	TransFormer	73.89	73.93	66.64	64.75
A + T + V*	NodeFormer	73.07	73.16	67.28	65.58
A* + T + V	NodeFormer	73.51	73.56	67.25	**66.15**
A + T* + V	NodeFormer	**74.24**	**74.20**	**67.86**	66.09

*This modality is the main modality in the NodeFormer

of Transformer encoders, the quality of the features can be further refined and improved, reducing redundancy and promoting complementarity between the modalities. Stacking multiple Transformer encoders can extract features at different abstraction levels. The lower-level encoders perform initial feature extraction on the input, while the higher-level encoders produce more abstract emotion representations. Such a wide spectrum of feature representations can better capture emotional information in different modalities and improve ERC performance.

6.3. Effect of Residual Connections

Table 5 shows the effect of residual connections [23] on the NodeFormer. Residual connections preserve the original input by adding it to the output of subsequent layers. In this way, NodeFormer retains the original information. When these connections are removed, the performance drops significantly, suggesting that adding the output $\boldsymbol{M}$ from the main modality and the output from the previous encoder $\boldsymbol{F}_{n-1}$ to the output of the

Table 4: *Impact of varying the number of Transformer encoders in NodeFormer on ERC performance.*

Number of Transformer Encoders n	**IEMOCAP**		**MELD**	
	Acc	**w-F1**	**Acc**	**w-F1**
0	72.81	72.87	62.98	61.16
1	72.69	72.78	64.13	62.18
2	73.81	73.82	67.44	65.85
3	73.82	73.88	**67.86**	**66.09**
4	**74.24**	**74.20**	67.10	65.64
5	73.95	74.03	67.29	65.51

Transformer encoder can improve the stability of deep feature extraction and ensure the interaction between the modalities.

Table 5: *Impact of residual connections in NodeFormer on ERC performance.*

Residual Connection		**IEMOCAP**		**MELD**	
Previous Output $\boldsymbol{F}_{n-1}$	**Main Input $\boldsymbol{M}$**	**Acc**	**w-F1**	**Acc**	**w-F1**
×	×	72.38	72.51	66.18	64.50
×	✓	72.75	72.79	66.94	65.19
✓	×	73.19	73.28	67.37	65.65
✓	✓	**74.24**	**74.20**	**67.86**	**66.09**

7. Conclusions and Future Work

In this paper, we propose a Multimodal Fusion Network called MM-NodeFormer to effectively capture multimodal emotion information for multimodal ERC. A NodeFormer combines characteristics of multimodal features and enhances the complementarity among multiple modalities. Experimental results show that MM-NodeFormer consistently outperforms the baselines, and the ablation studies validate the effectiveness of each modules in MM-NodeFormer. However, genders (same or mixed) in a multi-party conversation may also affect the emotional states of the utterances when the dialog evolves. In future work, we plan to add gender recognition as an auxiliary task to help the model determine the emotion of speakers, thereby considering gender factors in the emotion prediction task.

8. References

[1] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[2] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2594–2604.

[3] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations." in *Proc. IJCAI*, 2019, pp. 5415–5421.

[4] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (volume 1: Long papers)*, 2017, pp. 873–883.

[5] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, 2018, p. 2122.

[6] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," *arXiv preprint arXiv:1908.11540*, 2019.

[7] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies," *APSIPA Transactions on Signal and Information Processing*, vol. 3, p. e12, 2014.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. International Conference on Machine Learning*, 2021, pp. 8748–8763.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[11] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6818–6825.

[12] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," *arXiv preprint arXiv:2107.06779*, 2021.

[13] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7037–7041.

[14] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2FNET: Multi-modal fusion network for emotion recognition in conversation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4652–4661.

[15] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UNIMSE: Towards unified multimodal sentiment analysis and emotion recognition," *arXiv preprint arXiv:2211.11256*, 2022.

[16] Y. Liu, J. Li, X. Wang, and Z. Zeng, "EmotionIC: Emotional inertia and contagion-driven dependency modelling for emotion recognition in conversation," *arXiv e-prints*, pp. arXiv–2303, 2023.

[17] J. Li, Y. Liu, X. Wang, and Z. Zeng, "CFN-ESA: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition," *arXiv preprint arXiv:2307.15432*, 2023.

[18] Z. Li, F. Tang, M. Zhao, and Y. Zhu, "EmoCaps: Emotion capsule based model for conversational emotion recognition," *arXiv preprint arXiv:2203.13504*, 2022.

[19] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[20] Y. Mao, G. Liu, X. Wang, W. Gao, and X. Li, "DialogueTRM: Exploring multi-modal emotional dynamics in a conversation," in *Proc. Findings of the Association for Computational Linguistics: EMNLP*, 2021, pp. 2694–2704.

[21] L. Yuan, G. Huang, F. Li, X. Yuan, C.-M. Pun, and G. Zhong, "RBA-GCN: Relational bilevel aggregation graph convolutional network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2325–2337, 2023.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[24] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[25] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[26] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, L.-W. Ku *et al.*, "EmotionLines: An emotion corpus of multi-party conversations," *arXiv preprint arXiv:1802.08379*, 2018.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.