CMCF-SRNet: A Cross-Modality Context Fusion and Semantic Refinement Network for Emotion Recognition in Conversation

Xiaoheng Zhang
Beihang University
xiaoheng_zhang@buaa.edu.cn

Yang Li *
Beihang University
liyang@buaa.edu.cn

Abstract

Emotion recognition in conversation (ERC) has attracted enormous attention for its applications in empathetic dialogue systems. However, most previous researches simply concatenate multimodal representations, leading to an accumulation of redundant information and a limited context interaction between modalities. Furthermore, they only consider simple contextual features ignoring semantic clues, resulting in an insufficient capture of the semantic coherence and consistency in conversations. To address these limitations, we propose a cross-modality context fusion and semantic refinement network (CMCF-SRNet). Specifically, we first design a cross-modal locality-constrained transformer to explore the multimodal interaction. Second, we investigate a graph-based semantic refinement transformer, which solves the limitation of insufficient semantic relationship information between utterances. Extensive experiments on two public benchmark datasets show the effectiveness of our proposed method compared with other state-of-the-art methods, indicating its potential application in emotion recognition. Our model is available at https://github.com/zxiaohen/CMCF-SRNet.

1 Introduction

Emotion recognition in conversation (ERC) plays an important role in affective dialogue systems, aiming to understand and generate empathetic responses (Raamkumar and Yang, 2022). Most studies on ERC focus primarily on the textual modality (Majumder et al., 2019). Although they can be easily extended to multimodal paradigms by performing early or late fusion (Poria et al., 2017a), it is difficult to capture contextual interactions between modalities, which limits the utilization of multiple modalities. For instance, a tensor fusion network based on the utterance-level explicit alignment learning both intra-modality and inter-modality interactions via the Cartesian product was firstly created (Zadeh et al., 2017), then a low-rank multimodal fusion network to improve efficiency and reduce trainable parameters was designed (Liu et al., 2018). A conversational memory network aligns features from different modalities by fusing multi-view information (Hazarika et al., 2018b). In addition, a cross-modal transformer was integrated that learns attention between two-modal features, thus enabling implicit enhancement of the target modality (Tsai et al., 2019). A multimodal fusion graph convolutional network for ERC was put forward discussing the impact of fusion methods of various modalities (Hu et al., 2021).

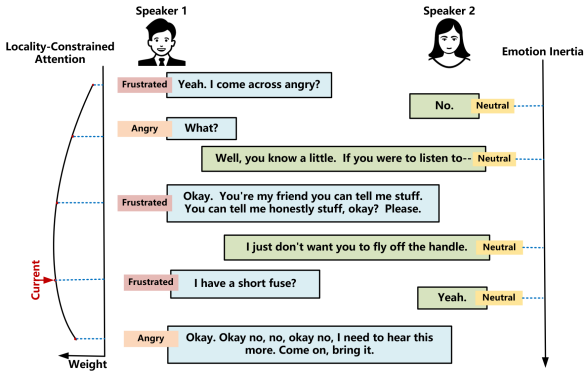

Fig. 1. An example conversation between two speakers with corresponding emotions evoked for each utterance illustrating the importance of local context.

However, these methods mostly use a simple concatenation ignoring complex interactions between modalities, resulting in leveraging context information insufficiently or the problem of data sparseness. Besides, they simply consider the emotional impact of context in the whole conversation but neglect the emotional inertia of speakers and the fact that the local context may have a higher impact than long-distance utterances.

As both the current utterance and the surrounding contexts are vital for the emotion perception, previous works proposed different methods including RNN-based models and graph-based models to explore contextual clues: A LSTM-based

*Corresponding author

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 13099–13110

model (Poria et al., 2017b) and an interactive conversational memory network ICON (Hazarika et al., 2018a) capture interaction and history context, while DialogueRNN model (Majumder et al., 2019) leverages distinct GRUs to capture speakers' contextual information.

Other popular approaches use graph-based neural networks and solve the context propagation issues in RNN-based architectures, including DialogueGCN (Ghosal et al., 2019) which first constructed the graph considering both speaker and conversation sequential information. Recent approaches like DAG-ERC (Shen et al., 2021) combined the advantages of conventional graph-based models and RNN-based models, a semantics GAT was employed to adjust the weight of knowledge (Tu et al., 2022), latent correlations have been leveraged among the utterances through a multi-branch graph network (Ren et al., 2022). Meanwhile, existing GNN models use different aggregation schemes for a node to aggregate feature messages from its neighbors (Yuan et al., 2022): Graph convolutional networks use mean pooling while graph attention networks aggregate neighborhood information with trainable attention weights to capture local details (Isufi et al., 2022). Furthermore, graph networks consider global graph information during aggregation and have been used to explore the semantic relationship between regional objects and global concepts (Zhu et al., 2022).

However, the existing graph-based methods also have limitations. First, they mostly ignore the semantic similarity between context utterances leading to a lack of semantic correlation. Second, these models learn node embeddings by capturing local network structure but ignore the position of the node within a broader context of the graph structure and the deep semantic features from a global view. To address this issue, we investigate a semantic graph-based transformer.

In this work, we propose a cross-modality context fusion and semantic refinement network (CMCF-SRNet). First, we investigate a cross-modality context fusion module to integrate textual and audio information considering the impact of the local context and the emotional inertia of speakers, achieved by a cross-modal locality-constrained attention. Second, we design a semantic refinement module to extract effective semantic features and contextual information including the nearby surroundings and distant information. The main contributions can be summarized as follows:

- Our proposed CMCF-SRNet is developed firstly by exploring cross-modal locality-constrained transformer to facilitate multimodal context fusion, bridging the gap of current works on ERC.
- We define a semantic graph to model the relations between neighboring utterances and a semantic graph-based transformer encoder is adopted to capture the global underlying semantic.
- We systemically analyze the importance of each component, including cross-modal transformer-based fusion, and semantic refinement methods. Experimental results demonstrate the performance of our proposed model.

2 Methodology

Given a dialogue $D = \{u_1, u_2, ..., u_N\}$, where N denotes the number of utterances and u_m is the m^{th} utterance in the conversation, the emotion recognition in conversation task aims to predict the emotion label for each utterance in the conversation. Each utterance involves two sources of data corresponding to acoustic (a) and textual (t) modalities represented as $u_m = \{u_m^a, u_m^t\}$ where $u_m^{(a)} \in \mathbb{R}^{d_a}$ for audio and $u_m^{(t)} \in \mathbb{R}^{d_t}$ for text, where d_a, d_t represent feature dimensions. The combined input features matrix for all utterances in a dialogue is given by: $X_i = [u_1^{(i)}, u_2^{(i)}, ...u_n^{(i)}]$ where $i \in \{a, t\}$.

The overall architecture of our proposed CMCF-SRNet is outlined in Fig. 2 and summarized as follows: (1) The acoustic/textual feature matrix for utterances is first fed to acoustic/linguistic embedding block to obtain unimodal representations, and then cross-modal locality-constrained attention (LCA) is utilized to generate high-level cross-modal features which go into an attentive selection block; (2) We define a semantic graph and employ the relational graph convolutional network to capture the inter-utterance dependence, then leverage an aggregation of effective semantic features by integrating a semantic-position encoding; (3) The nodes embeddings are fed into the classifier to obtain the final prediction. In the following three subsections, we discuss in detail the specific implementation of the proposed innovation modules.

2.1 Cross-modality Context Fusion Module

We consider the order of the utterances by adding the triangle positional embedding (PE) directly to X_i while $i \in \{a, t\}$, and we define the query $Q_i^{(h)}$,

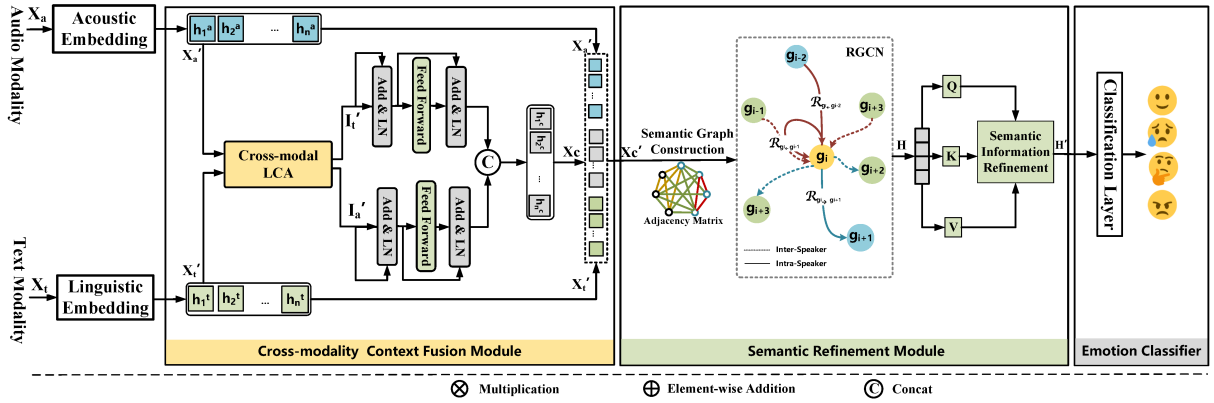

Fig. 2. Illustration of the proposed CMCF-SRNet consisting of two modules: cross-modal context fusion module and semantic refinement module (LCA: locality-constrained attention).

the key $K_i^{(h)}$ and the value $V_i^{(h)}$ vector for encoding input features $X_i \in \mathbb{R}^{n \times d_i}$ as shown in Fig. 3 (a). An attention map of attention weights for a single attention head $\alpha^{(h)} \in \mathbb{R}^{n \times n}$ is obtained by the attention mechanism and is used to compute a weighted sum of the values and obtain the output.

$$O_i^{(h)} = \text{softmax}(\frac{Q_i^{(h)}(K_i^{(h)})^T}{\sqrt{k}})V_i^{(h)} \quad (1)$$

$$\hat{O}_i{}^{(h)} = [O_i^{(1)} \oplus O_i^{(2)} \oplus \cdots \oplus O_i^{(N)}]W \quad (2)$$

where $W \in \mathbb{R}^{kN \times d_i}$, N represents the total number of heads. Finally, we add a residual connection followed by a layer norm and obtain X_i'.

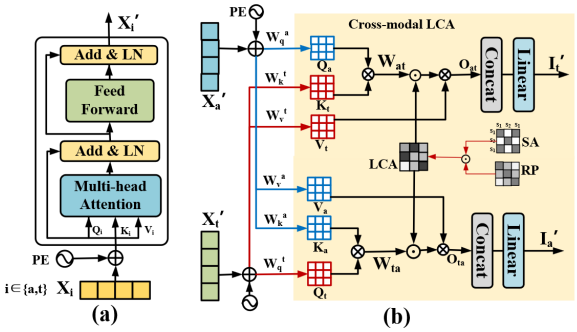

Fig. 3. (a) Unimodal embedding. (b) Cross-modal LCA.

After an intra-modal transformer to capture the global temporal dependencies of unimodal features, we apply a cross-modal locality-constrained transformer to capture the local contextual information focusing on correspondences between different modalities. We extend the traditional transformer to a two-stream cross-modal transformer to model interactions between two modalities, where each cross-modal transformer block is combined with a cross-modal locality-constrained attention layer. The attention layer could combine the information from the different sources of data to transform the text features using the feature map of audio. Querys, Keys, and Values has been defined as $Q_i^{(h)} = X_i' W_{h,q}$, $K_j^{(h)} = X_j' W_{h,k}$, $V_j^{(h)} = X_j' W_{h,v}$ for a single attention head where $i, j \in \{a, t\}, i \neq j$. Considering that to predict the emotion of an utterance, the speaker's recently stated utterance has the greatest correlation with its emotion, thus, we propose a locality-constrained and speaker aware attention LCA (Fig. 3 (b)) by masking the traditional weight map $W_{ij} = softmax(Q_i K_j^T)$ in Eq. (3) as in Eq. (4).

$$O_{ij}(Q, K) = W_{ij} \cdot V_j \quad (3)$$

$$O_{ij}(Q, K) = (W_{ij} \odot LCA) \cdot V_j \quad (4)$$

We design intra-speaker masks SA to focus on the utterances of the current speaker and model the emotional inertia of this interlocutor's emotional flow on the current utterance:

$$\text{SA}_{m,n} = \begin{cases} 1 & \text{if } s_m = s_n; \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

where s_m, s_n are respectively the speakers of utterances u_m and u_n. As the emotion of current utterance is more affected by the local utterances close to it, a common idea is to apply a fixed window, but in order to solve the problem that the fixed-window method treats utterances in the window equally, we calculate the relative position weighting RP of h_m and h_n, then feed into a sigmoid function. Finally, we apply an element-wise product to obtain $LCA = \text{sigmoid}(RP) \times SA$, which combines both local context and speaker information.

$$\text{RP}_{m,n} = \begin{cases} \text{M} - C(n-m)^2 & \text{if } m, n \leq N; \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

where N is the actual length of the dialogue, and both M and C are hyperparameters. Here, we set M and C to 5 and 1.5 respectively.

To obtain the fusion representation combing both intra-modal and cross-modal contextual information from two modalities, an attentive selection block is proposed to distribute different importance to different modalities, we propose a model-

level fusion strategy instead of a simple concatenation (Chen and Jin, 2016). Experimental results in Section V verify the effectiveness.

We extract utterance-level acoustic features $h_m^{(a)} \in \mathbb{R}^d$, text features $h_m^{(t)} \in \mathbb{R}^d$, cross-modal features $h_m^{(c)} \in \mathbb{R}^d$ for each utterance u_m (where u_m is the m^{th} utterance in the conversation). Then we equalize feature dimensions of all inputs and concatenate them together considering different contributions of different modalities to focus on important modalities. Technically, at a given time, given the input feature $H = [H^{(1)}, H^{(2)}, ..., H^{(K)}]$ with K the number of modalities. The score for each modality is computed by:

$$a_i = \mathrm{ReLU}(W^T H^{(i)} + b) \tag{7}$$

$$\alpha_i = \frac{exp(a_i)}{\sum_{j=1}^{K} exp(a_j)} \tag{8}$$

the attention scores $\alpha_{att} \in R^{1\times K}$ where $K = 3$. The final multimodal features $g_m \in \mathbb{R}^{d_2}$ are generated as follows with the output $X_c' \in \mathbb{R}^{n\times d_2}$:

$$g^{(j)} = concat([\alpha_1 H^{(1)}, ..., \alpha_K H^{(K)}]) \tag{9}$$

2.2 Semantic Refinement Module

To explore the semantic relationships between utterances in a dialogue, a novel model for semantic information refining is proposed, which is illustrated in the semantic refinement module in Fig. 2. It mainly consists of two stages: relational semantic graph construction and semantic information refinement. The well-defined semantic graph is fed into a two-layer RGCN to compute semantic features of utterances and their interaction relations. Then the global semantic information is further extracted by a semantic graph-transformer.

Semantic Graph Construction: To establish semantic relations between the nearby utterances and to capture both inter-speaker and intra-speaker effects, we define a semantic graph G_s= (V_s,E_s) based on the conversational semantic-aware dependency. Each utterance is represented by a node and different connection edges represent directed relations (past and future), V_s denotes a set of utterance nodes, and $E_s \subset V_s \times V_s$ is a set of relations that represent the semantic similarity between the utterances, defined as Eq. (10).

$$sim_{i,j} = 1 - \arccos(\frac{g_i^T g_j}{\|g_i\|\|g_j\|}) \tag{10}$$

We define intra-relations between the utterances spoken by the same speaker $R_{intra} \in \{U^{S_i} \to U^{S_i}\}$ and inter-relations by different speakers, $R_{inter} \in \{U^{S_i} \to U^{S_j}\}_{i\neq j}$. We further consider a context window using $\mathcal{P}$ and $\mathcal{F}$ as hyperparameters to denote relations between the past $\mathcal{P}$ utterances and future $\mathcal{F}$ utterances for every utterance. The relational semantic graph can be regarded as a local-view modeling of the relationships between utterances in a dialogue and covering semantics features.

Semantic Information Refinement: In this paper, a modified relational graph convolution layer is adopted to capture local dependency defined by the relations. The node representations and edge weights are feed into a two-layer correlation-based RGCN which can be summarized as follows, here, we introduce the concept of aggregate functions to generalize the above mechanism:

$$h_i^{(1)} = \sigma(\sum_{r\in\mathcal{R}} \sum_{j\in N_i^r} \frac{a_{i,j}}{q_{i,r}} W_r^{(1)} g_j + a_{i,i} W_0^{(1)} g_i)$$

$$h_i^{(2)} = \sigma(\sum_{j\in N_i^r} W^{(2)} h_j^{(1)} + a_{i,i} W_0^{(2)} g_i) \tag{11}$$

where N_i^r denotes the neighboring indices of each node under relation $r \in \mathcal{R}$, $W_i^{(m)}$ are learnable parameters, $\sigma(\cdot)$ is the activation function as ReLU. In this way, each graph convolution layer models the interaction between utterances, and refines the semantic features.

Then, we adopt a semantic graph-transformer to extract global semantic information from the node feature taking in consideration the relative position of utterances (Fig. 4). It adopts the vanilla multi-head attention into graph learning by taking into account nodes connected via edges. Given node features $H = [h_1, h_2, ..., h_n]$ obtained from RGCN, we define two encodings to represent semantic relationship between two nodes in a graph. The first is relative position encoding $\mathcal{P}$, each vector of $\mathcal{P}$ represents the topological relation represented by their shortest path distance between two nodes, the second is semantic encoding $\mathcal{S}$ defined by Eq. (10), we take an addition operation and obtain $\mathcal{SP}$.

$$a_{ij} = \frac{(W_q h_i)^T (W_k h_j)}{\sqrt{d}^{value}} + \Phi_{ij}^{sem} \tag{12}$$

$$\Phi_{ij}^{sem} = q_i \mathcal{SP}_{\phi_{ij}^{sem}} + k_j \mathcal{SP}_{\phi_{ij}^{sem}} \tag{13}$$

$$h_i' = \sum_{i=1}^{N} \hat{a}_{ij} (v_j + \mathcal{SP}_{\phi_{ij}^{sem}}) \tag{14}$$

Previous methods focus on encoding graph information into either the attention map or input features. First, our method encodes positional and semantic information represented by edge weight

into attention map to take the global context structure into consideration. Moreover, it encodes the hidden features of value as shown in Eq. (14).

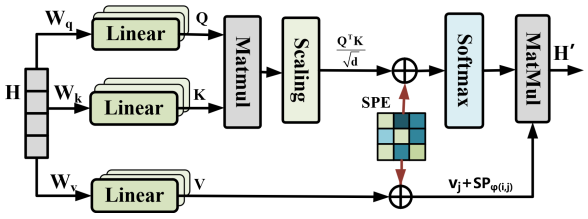

Fig. 4. Semantic Information Refinement.

2.3 Emotion Classifier

The output of graph transformer is fed into a MLP with fully connected layers and get the prediction values of the utterance u_i under each emotion label:

$$h_i = \mathrm{ReLU}(W_1 h_i' + b_1) \quad (15)$$

$$\mathcal{P}_i = softmax(W_2 h_i + b_2) \quad (16)$$

$$\hat{y}_i = \mathrm{argmax}(\mathcal{P}_i) \quad (17)$$

where $\hat{y}_i$ is the emotion label predicted for the utterance u_i. We choose the categorical cross-entropy loss function during training as is shown below:

$$\mathcal{L} = -\frac{1}{\sum_{i=1}^{N} L_i} \sum_{n=1}^{N} \sum_{i=1}^{C} y_i \cdot log\hat{y}_i \quad (18)$$

where N is the number of conversations and L_i is the number of utterances in the i_{th} conversation.

3 Experiments and Results

3.1 Datasets

In this section, we conduct several experiments to evaluate our proposed method and compare it with state-of-the-art baselines on two benchmark datasets, the dataset statistics are given in Table 1:

Table 1: Statistics of IEMOCAP and MELD datasets.

Statistics	IEMOCAP			MELD		
	train	valid	test	train	valid	test
Nb of dialogues	120		31	1039	114	280
Nb of utterances	4290/5810		1241/1623	9989	1109	2610

- **IEMOCAP** (Busso et al., 2008) dataset contains approximately 12 hours of dyadic emotional improvised and scripted conversations (10039 utterances). The labelling of each utterance was determined by 3 annotators as the following categorical labels: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise. To compare with state-of-the-art frameworks, we adopt their dataset settings respectively the first four categories for the 4-way condition (Lian et al., 2021) and the first six categories for the 6-way conditions. Following previous works, utterances from the first 8 speakers are used as the training and validation sets while the others are used as the testing set.
- **MELD** (Poria et al., 2019) is a large-scale multi-party conversational dataset which contains 13708 utterances and 1433 conversations from TV series *Friends*, and each utterance is annotated with one of the following labels: anger, joy, sadness, neutral, disgust, fear and surprise.

3.2 Implementation Details and Metrics

We performed all experiments on the Pytorch deep learning framework with the Intel Core i7-12700H and the NVIDIA RTX3060 GPU. The software environment includes Python 3.9, Pytorch 1.12.1, and CUDA 11.3. Adam optimizer with an initial learning rate of 0.0001 is used to optimize the parameters in the proposed CMCF-SRNet and a dropout rate of 0.5 is adopted. The head number is set to 4 for cross-modal transformer and 2 for graph-transformer. Besides, audio features (size 100) are extracted using OpenSmile (Eyben et al., 2010) and text features (size 768) are extracted using sBERT (Reimers and Gurevych, 2019). We re-run on each dataset five times and calculate the mean and standard deviations.

We evaluate the performance of emotion recognition using the following as evaluation metrics: **WAA** is a weighted average accuracy over different emotion classes with weights proportional to the number of utterances in a class. **WF1** is a weighted mean F1 over different emotion categories with weights proportional to the number of utterances in a particular class.

$$WAA = \frac{\sum_{j=1}^{C} N_j * Accuracy_j}{\sum_{j=1}^{C} N_j} \quad (19)$$

$$WF1 = \frac{\sum_{j=1}^{C} N_j * F1_j}{\sum_{j=1}^{C} N_j} \quad (20)$$

3.3 Overall Performance

For comparison, we implement following state-of-the-art baseline approaches to evaluate the performance of our proposed method:

BC-LSTM (Poria et al., 2017c) uses a bi-directional LSTM to encode contextual information, but ignoring the speaker-specific information.

DialogueGCN (Ghosal et al., 2019) is the first to model a conversation by a graph, transforms

Table 2: Results on IEMOCAP (6-way) and MELD (⋆ represents models with multimodal (A+T+V) setting).

Models	Year	IEMOCAP(6-way): Emotion Categories								MELD
		Happy	Sad	Neutral	Angry	Excited	Frustrated	Average		Average
		WF1(%)	WF1(%)	WF1(%)	WF1(%)	WF1(%)	WF1(%)	WAA(%)	WF1(%)	WF1(%)
Bc-LSTM	2017c	35.6	69.2	53.5	66.3	61.1	62.4	59.8	59.0	50.8
DialogueGCN	2019	42.7	**84.5**	63.5	64.1	63.0	**66.9**	65.2	64.1	55.8
CTNet⋆	2021	51.3	79.9	65.8	67.2	**78.7**	58.8	68.0	67.5	60.5
A-DMN⋆	2022	50.6	76.8	62.9	56.5	77.9	55.7	64.6	64.3	60.4
I-GCN⋆	2022	50.0	83.8	59.3	64.6	74.3	59.0	65.5	65.4	60.8
MMDFN⋆	2022	42.2	78.9	66.4	69.7	75.5	66.3	68.2	68.1	59.4
CMCF-SRNet (Ours)	2023	**52.2**±0.5	80.9±0.2	**68.8**±0.5	**70.3**±0.6	76.7±0.3	61.6±0.7	**70.5**±0.8	**69.6**±0.7	**62.3**±0.6

the emotion classification into a graph-based node classification problem.

MMGCN (Hu et al., 2021) uses multimodal dependencies and speaker information effectively and applies GCN to obtain contextual information.

CTNet (Lian et al., 2021) utilizes transformer to obtain the multimodal rerpesentation by modeling the intra-modal and cross-modal interactions.

A-DMN (Xing et al., 2022) models self and inter-speaker influences and then synthesizes this two factors to update the memory.

I-GCN (Nie et al., 2022) utilize the graph structure to represent conversation at different times and apply the incremental graph structure to imitate the process of dynamic conversation.

MMDFN (Hu et al., 2022) proposes a graph model where both speaker dependency of the interlocutors is leveraged and latent correlations are captured.

To verify the effectiveness of our proposed method, we compare our proposed CMCF-SRNet with state-of-the-art baseline approaches on the IEMOCAP (4-way), IEMOCAP (6-way) and MELD datasets on the overall performance and for each emotion category. As is shown in Table 2, our model outperforms all the baselines mentioned above on the two datasets. For the IEMOCAP (4-way) dataset, ours achieves the new state-of-the-art record, 86.5% on F1 and 86.9% on WAA, which shows an absolute improvement of 2.0% on F1 score. For the IEMOCAP (6-way) dataset, our proposed method also succeeds with 70.5% on WAA and 69.6% on F1 which outperforms Bc-LSTM and DialogueGCN by 10.7% on WAA, 10.6% on WF1 and 5.3% on WAA, 5.5% on WF1 possibly due to the cross-modal context fusion architecture applied in our proposed model; in addition, it outperforms CTNet and MMDFN which utilize multimodal fusion approaches by 2.3%~2.5% on WAA and 1.5%~2.1% on WF1, the reason lies in that these methods focus on the multimodal representation ignoring the semantic relationship between utterances. Our proposed CMCF-SRNet also outperforms I-GCN which highlights the semantic correlation information of utterances without considering multimodal fusion approach.

In addition, we present classification accuracies and F1 scores for each emotion category and visualize the confusion matrices of the testing set in Fig. 5. For the IEMOCAP (6-way) dataset, the improvements on classification performance can be seen for most emotion categories over existing approaches (Table 2). Specifically, we notice an improvement of F1-score for happy, neutral, angry and excited emotions which show the improved ability of the model to identify relevant emotions. Meanwhile, we find neutral and anger emotions can be confused with the frustration emotion (Fig. 5 (a)) as the majority of the utterances are labeled as the frustration. Also, the happiness emotion can be confused with the excitement emotion (Fig. 5 (a)) due to our similar perception of these emotions.

Table 3: Performance on IEMOCAP (4-way).

Methods	Year	IEMOCAP(4-way)	
		Modality	WF1(%)
Bc-LSTM	2017c	T	76.8
DialogueGCN	2019	T	81.7
CMCF-SRNet (Ours)	2023	T	**85.6**
CTNet	2021	A+T	83.6
COGMEN	2022	A+T+V	84.5
CMCF-SRNet (Ours)	2023	A+T	**86.5**

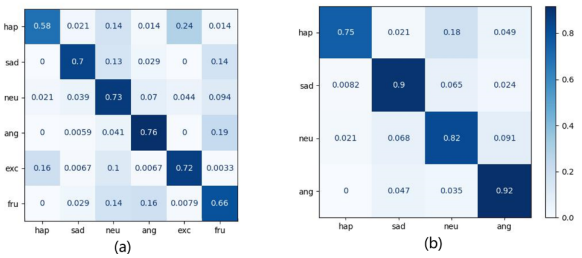

Fig. 5. Visualization the confusion matrices: (a) on the IEMOCAP (6-way); (b) on the IEMOCAP (4-way).

4 Discussion

4.1 Effect of Cross-modality Context Fusion

First, we conduct uni-modal experiments using text modality and our proposed method still gives comparable performance compared to the SOTA unimodal architectures (Table 3). As shown in Table 4, adding more information via other modalities helps to improve the performance.

To verify the effectiveness of our cross-modal locality constrained transformer-based contextual fusion strategy, we conduct the ablation experiments as listed in Table 4: 1) Without cross-modal Locality Constrained Attention (w/o LCA): We remove the transformers and combine utterance-level features directly with Attentive Selection Block; 2) Without Attentive Selection Block (w/o ASB); 3) Ours: Our proposed method. The results demonstrate that our proposed CMCF-SRNet with LCA significantly improved the WF1 and WA indexes. After adding LCA, the WA and WF1 of the model were improved by 3.2% and 3.4% respectively, indicating that the cross-modal transformer can comprehensively improve the performance. Then, as is shown in Fig. 6, we take the lexical modality for example and visualize its attention weights in conversations after different components. The red rectangles at the first line indicate that the 10th and the 14th utterances in the conversation show more importance for the emotion detection according to the intra-modal transforme while that in the second line indicates that according to the cross-modal transformer the 4th to 7th utterances should be paid more attention. These results verify that the outputs of cross-modal transformer contribute to conversational emotion recognition.

Fig. 6. Visualization using attention weights heatmap: (a) Intra-modal transformer; (b) Cross-modal LCA.

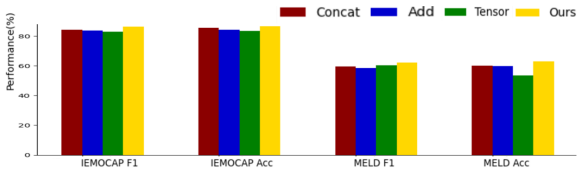

Fig. 7. Performance of different fusion strategies compared with ASB on MELD and IEMOCAP.

To verify the effectiveness of our attentive selection block (ASB), we implement three comparison methods. Experimental results (Table 4) demonstrate that our ASB achieves the best performance. It shows an absolute improvement over Add on WA by 2.5% on IEMOCAP, probably because Add copes with the multimodal features equally and cannot highlight emotion-relevant modalities while ASB can prioritize important modalities via the attention mechanism. Meanwhile, it also shows an improvement over Concatenate and Tensor Fusion which may suffer from the curse of dimensionality by 1.2% and 3.2% as our proposed method can generate more effective smaller-size multimodal features for emotion recognition.

Table 4: Comparison with unimodal architectures and ablation study on IEMOCAP(4-way) and MELD.

Methods	IEMOCAP (4-way)		MELD	
	WAA(%)	WF1(%)	WAA(%)	WF1(%)
T	85.6	85.1	60.4	59.7
A	60.6	59.2	55.5	53.2
A+T	**86.8**	**86.5**	**62.8**	**62.3**
w/o LCA	83.6	83.2	60.5	59.3
w/o ASB	84.5	84.1	61.1	60.3
w/o SEW	84.2	83.6	59.8	57.9
w/o SPE	83.6	83.8	60.8	59.6
Ours	**86.8**	**86.5**	**62.8**	**62.3**
Concatenate	85.6	84.2	60.2	59.62
Add	84.3	83.9	59.8	58.5
Tensor Fusion	83.6	83.1	53.5	60.3
Ours	**86.8**	**86.5**	**62.8**	**62.3**

4.2 Effect of Semantic Refinement

To observe the effect of the graph-based semantic refinement components, we visualize the features with and without the semantic refinement components (Fig. 8). We easily notice a better formation of emotion clusters proving the necessity of capturing semantic dependency in utterances. Additionally, we conduct ablation experiments on the correlation-based RGCN and Semantic Graph-Transformer respectively, specifically, we respectively remove the semantic edge weight (SEW) in the RGCN and semantic-positional encoding (SPE) in the Graph-Transformer, after removing the SEW, WA and WF1 on IEMOCAP decreased by 2.6%, 1.9% respectively, while after removing the SPE, WA and WF1 on IEMOCAP decreased by 3.2%, 2.7% respectively, which indicates that the proposed semantic encoder is necessary, the result in Table 4 shows the advantages of focusing on emotional semantic clues.

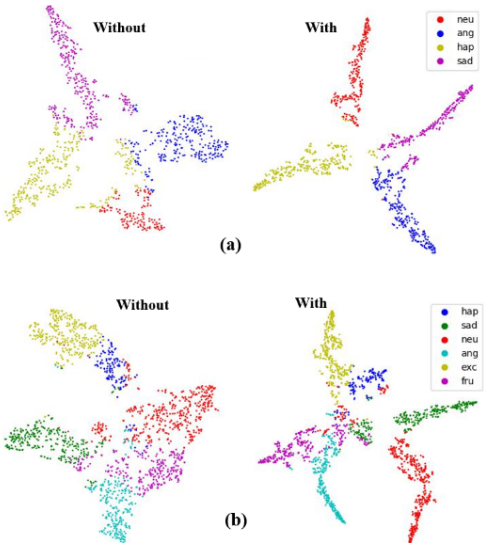

Fig. 8. T-SNE representation with and without semantic information refinement components respectively on (a) IEMOCAP (4-way) and (b) IEMOCAP (6-way).

4.3 Visualization and Interpretability

Given the importance of interpretability in machine learning, we investigate the necessity of local context realised by cross-modal LCA and global semantic context captured by semantic refinement module. We explore the distribution of distances between the target utterance and its second (2nd) highest attended utterance according to our attention scores for all the utterances correctly classified. First, most of the correctly classified utterances depend on their local context when a significant portion is also present for the distant context. Besides, the dependence on distant context shows more significance for the 2nd highest attention, which highlight the importance of the long-term emotional dependency. Meanwhile, the contextual dependence exists both towards the past and the future utterances.

Moreover, we conduct experiments with multiple window sizes as presented in Fig. 10. The window size can be modified during the training period. A larger window size would result in better performance for cases where the inter and intra speaker dependencies are maintained for longer sequences. In contrast, a smaller window size would be better where the topic frequently changes in dialogues and speakers are less affected by another speaker. These results support our design combining the locality-constrained attention and semantic refinement from a global-view.

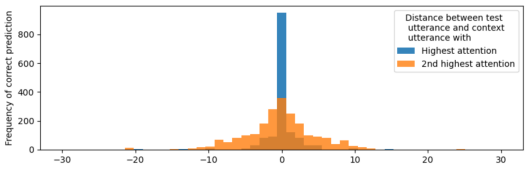

Fig. 9. Histogram of distance between the target utterance and its (2nd) highest attended utterance on MELD.

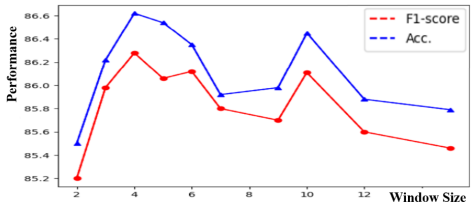

Fig. 10. Comparison for various window sizes.

5 Conclusion

In this paper, we propose a novel framework for multimodal emotion recognition which contains two innovative modules: The cross-modal locality-constrained context fusion leverages the transformer-based method to focus on localness, effectively improved the multimodal interaction. The semantic refinement module makes full use of the semantic relation information from a global view. Experiments on two public datasets and the results demonstrate that our proposed CMCF-SRNet is superior to the existing state-of-the-art methods. The ablation experiments prove the effectiveness of the two innovative modules. The detailed discussion shows that our proposed CMCF-SRNet has satisfactory generalization ability and interpretability, indicating that it has the potential for practical use for emotion recognition.

Limitation

Although experiments on two public datasets show the effectiveness of our proposed method compared with other state-of-the-art methods, we notice that our proposed model fails to distinguish similar emotions effectively going through the prediction results, as frustrated and anger, happy and excited (Fig. 5(a)). Moreover, our proposed model tends to misclassify samples of other emotions to neutral on MELD due to the majority proportion of neutral samples in these datasets. We will address this issue in future work by integrating a component for capturing the fine-grained emotions.

Acknowledgements

We appreciate the insightful suggestions from the anonymous reviewers to further improve our paper. This work was supported in part by the National Natural Science Foundation of China under Grant 62201023 and Beijing Natural Science Foundation under Grant Z220017, and in part by the Beijing Municipal Education Commission-Natural Science Foundation [KZ202110025036].

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Shizhe Chen and Qin Jin. 2016. Multi-modal conditional attention fusion for dimensional emotion prediction. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 571–575.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. 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)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. 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, Brussels, Belgium. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. 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, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7037–7041.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. 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 (Volume 1: Long Papers)*, pages 5666–5675, Online. Association for Computational Linguistics.

Elvin Isufi, Fernando Gama, and Alejandro Ribeiro. 2022. Edgenets: Edge varying graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7457–7473.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. 2022. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164, Seattle, United States. Association for Computational Linguistics.

Zheng Lian, Bin Liu, and Jianhua Tao. 2021. Ctnet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:985–1000.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Weizhi Nie, Rihao Chang, Minjie Ren, Yuting Su, and Anan Liu. 2022. I-gcn: Incremental graph convolution network for conversation emotion detection. *IEEE Transactions on Multimedia*, 24:4471–4481.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017a. 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)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017b. 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)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017c. 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)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Aravind Sesagiri Raamkumar and Yinping Yang. 2022. Empathetic conversational systems: A review of current advances, gaps, and opportunities. *IEEE Transactions on Affective Computing*, pages 1–20.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Minjie Ren, Xiangdong Huang, Wenhui Li, Dan Song, and Weizhi Nie. 2022. Lr-gcn: Latent relation-aware graph convolutional network for conversational emotion recognition. *IEEE Transactions on Multimedia*, 24:4422–4432.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. 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 (Volume 1: Long Papers)*, pages 1551–1560, Online. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Shaojie Bai, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Geng Tu, Bin Liang, Dazhi Jiang, and Ruifeng Xu. 2022. Sentiment- emotion- and context-guided knowledge selection framework for emotion recognition in conversations. *IEEE Transactions on Affective Computing*, pages 1–14.

Songlong Xing, Sijie Mai, and Haifeng Hu. 2022. Adapted dynamic memory network for emotion recognition in conversation. *IEEE Transactions on Affective Computing*, 13(3):1426–1439.

Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2022. Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–19.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, and Xiao Xiao. 2022. Multimodal emotion classification with multi-level semantic reasoning network. *IEEE Transactions on Multimedia*, pages 1–13.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section IV

☑ A2. Did you discuss any potential risks of your work?
Section IV

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section I

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section III

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section III

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section III

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section IV

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section III

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.