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

Code:None

—— ACL 2023

2023.10.29 • ChongQing

Leibniz-Institut für Sozialwissenschaften

Reported by Tingting Zhang

Motivation

(1) Many methods mostly use a simple concatenation ignoring complex interactions between modalities, resulting in leveraging context information insufficiently or the problem of data sparseness.

(2) 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.

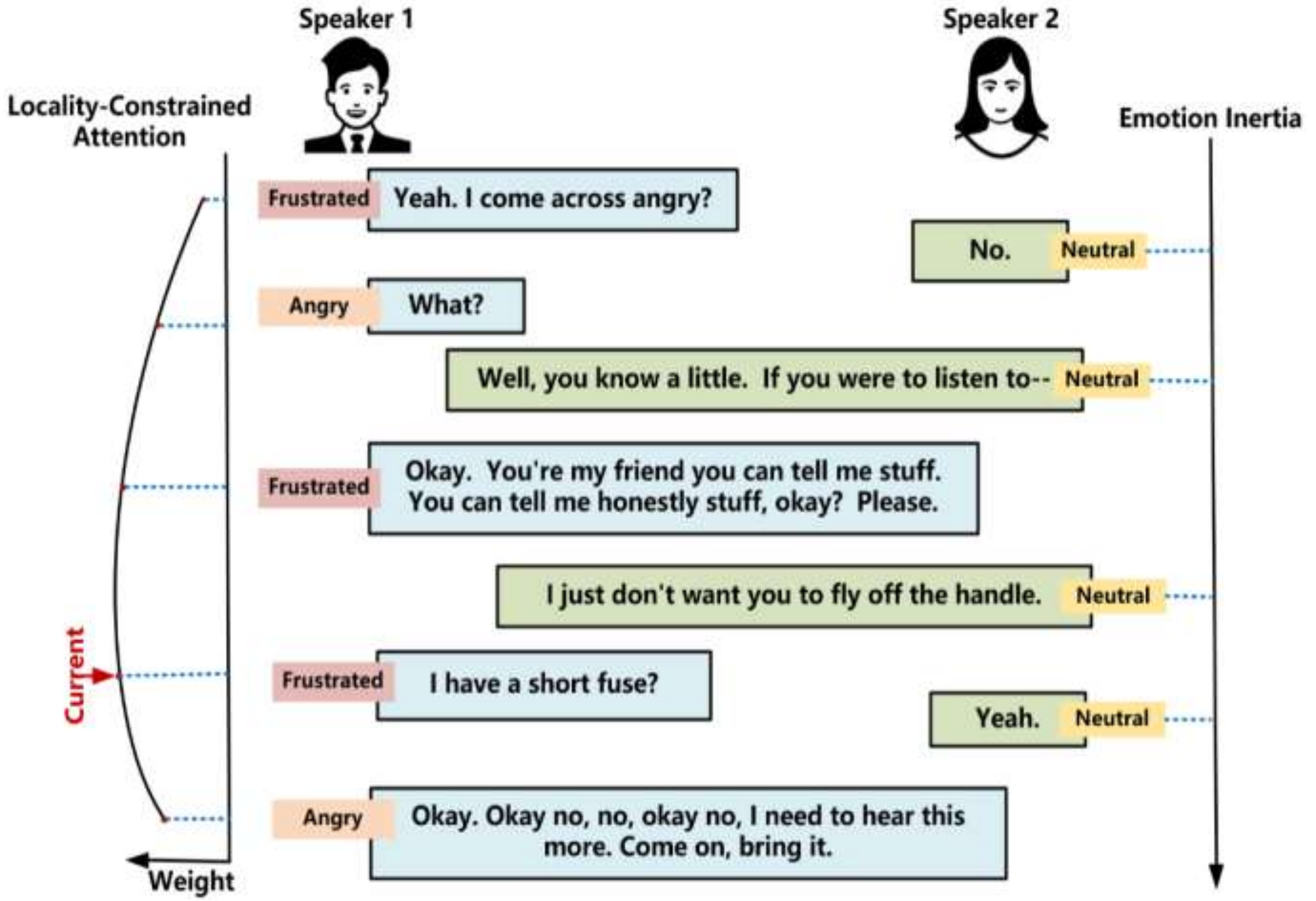


Fig. 1. An example conversation between two speakers with corresponding emotions evoked for each utterance illustrating the importance of local context.

Motivation

(3) 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.

Overview

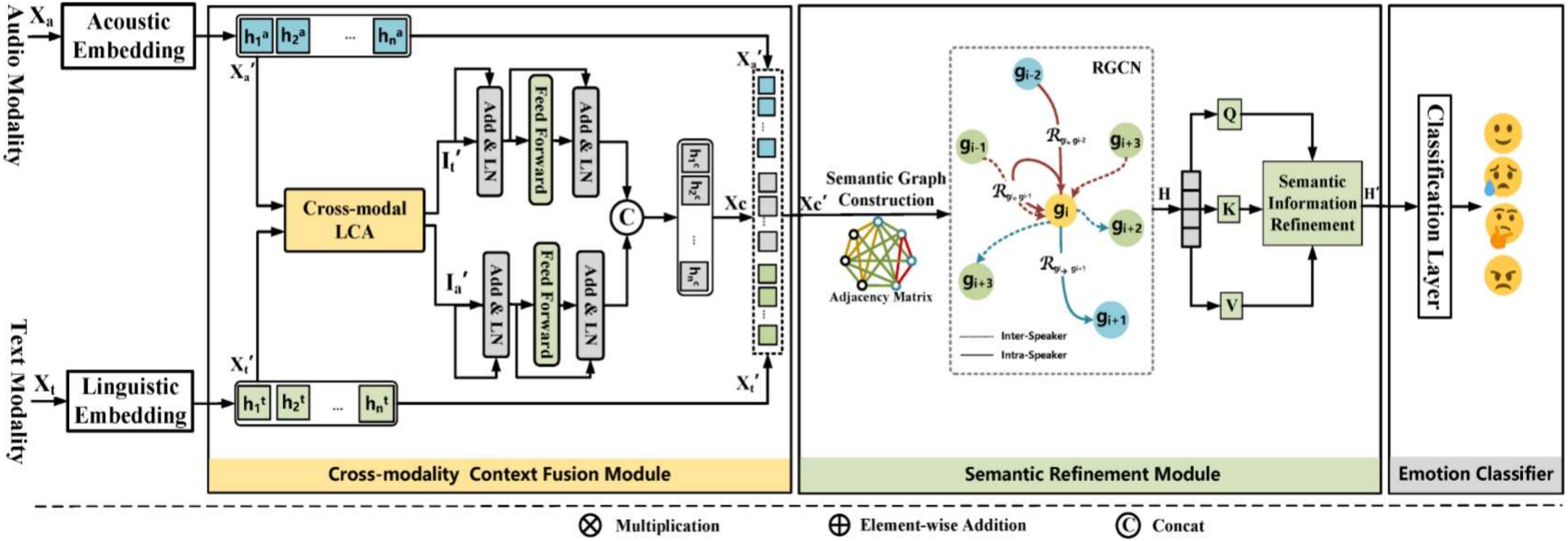


Fig. 2. Illustration of the proposed CMCF-SRNet consisting of two modules: cross-modal context fusion module and semantic refinement module (LCA: locality-constrained attention).

Method

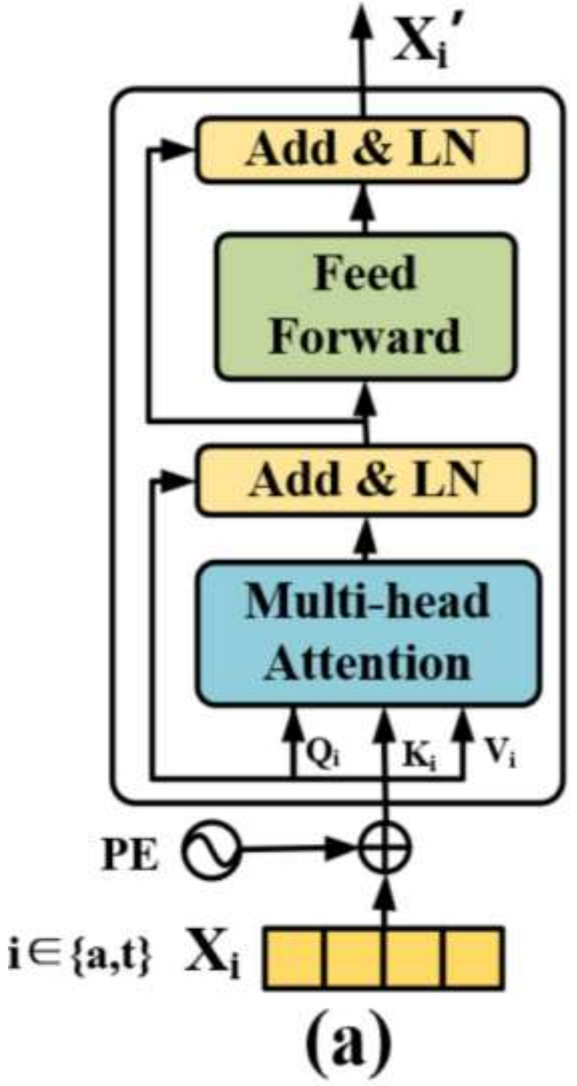
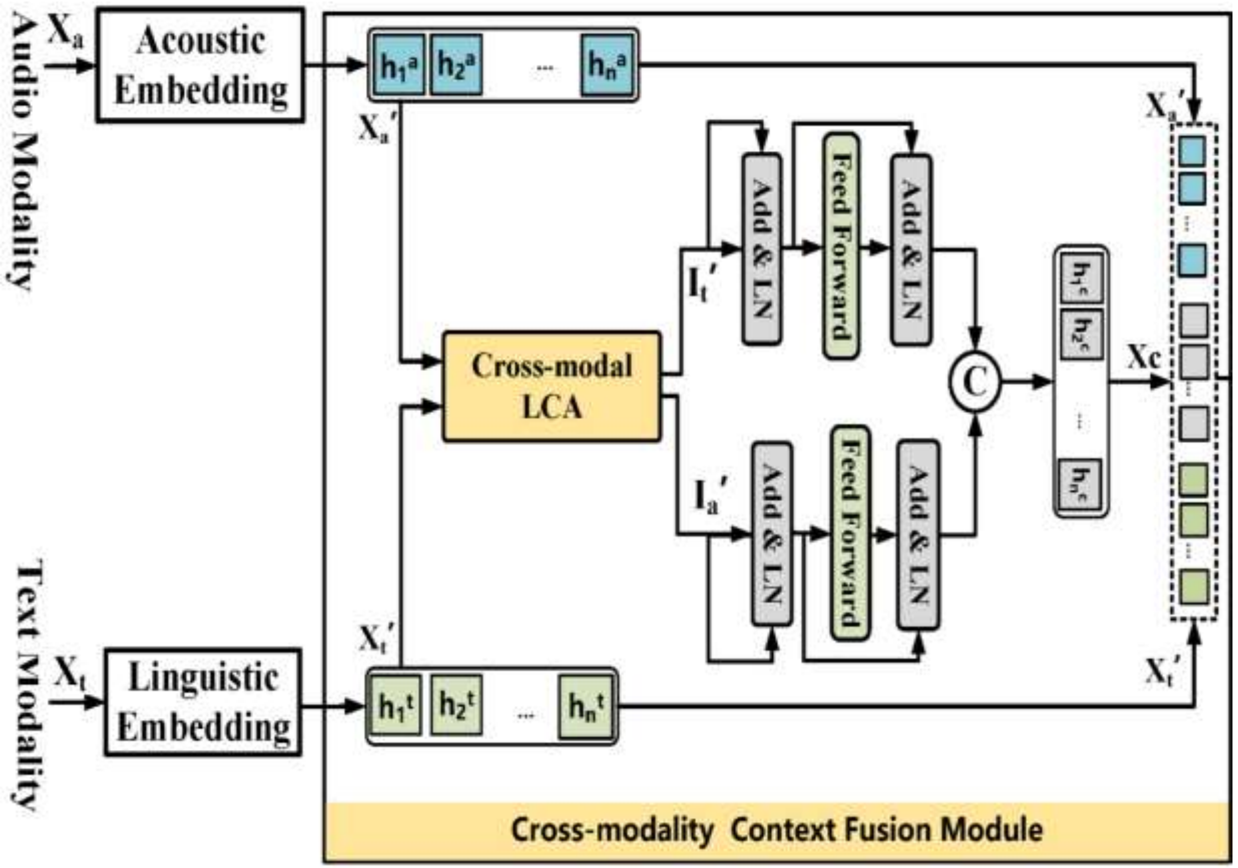


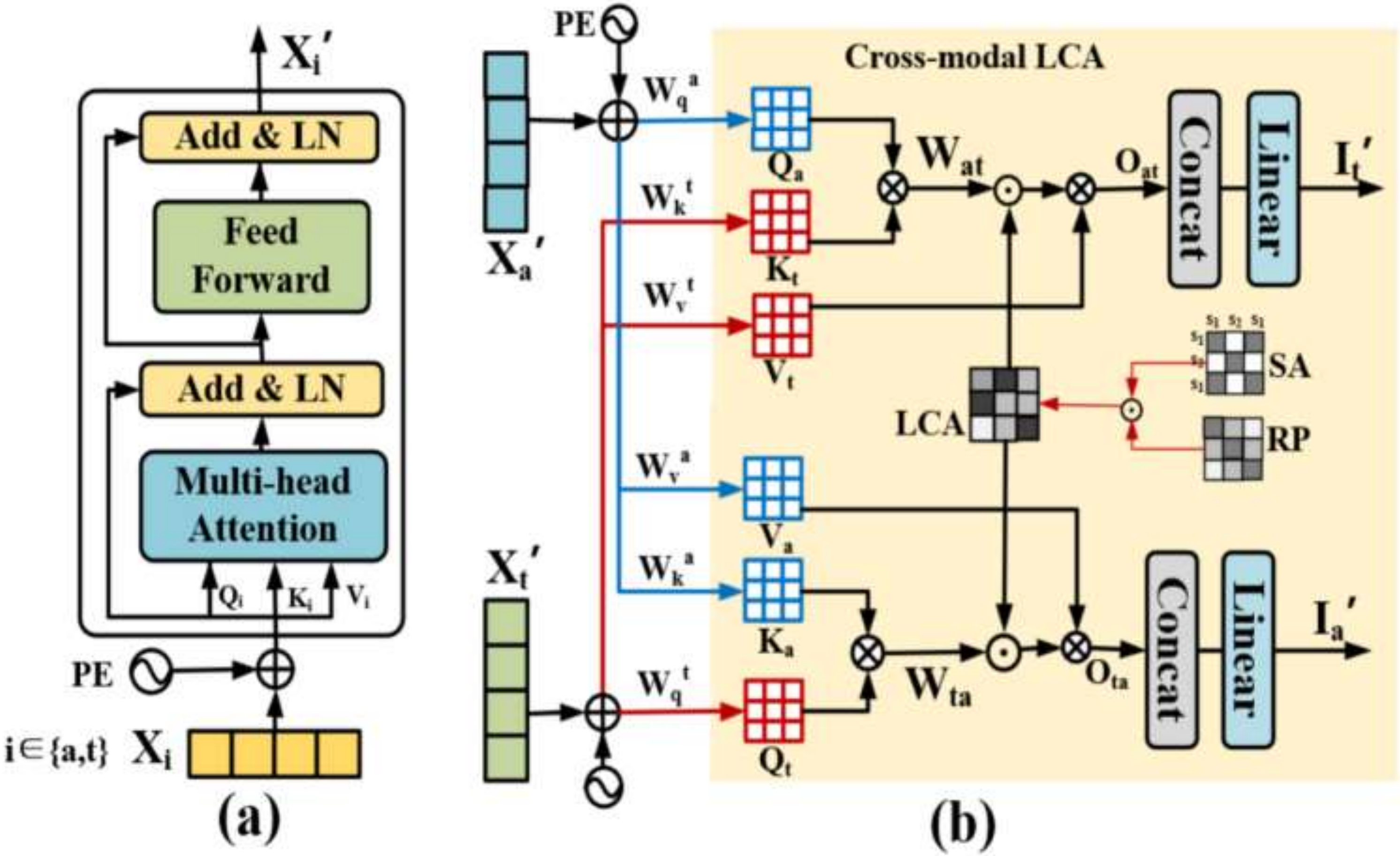
Fig. 3. (a) Unimodal embedding.

$$O_i^{(h)} = \mathrm{softmax}\left(\frac{Q_i^{(h)}(K_i^{(h)})^T}{\sqrt{k}}\right)V_i^{(h)} \tag{1}$$

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



Method



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

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

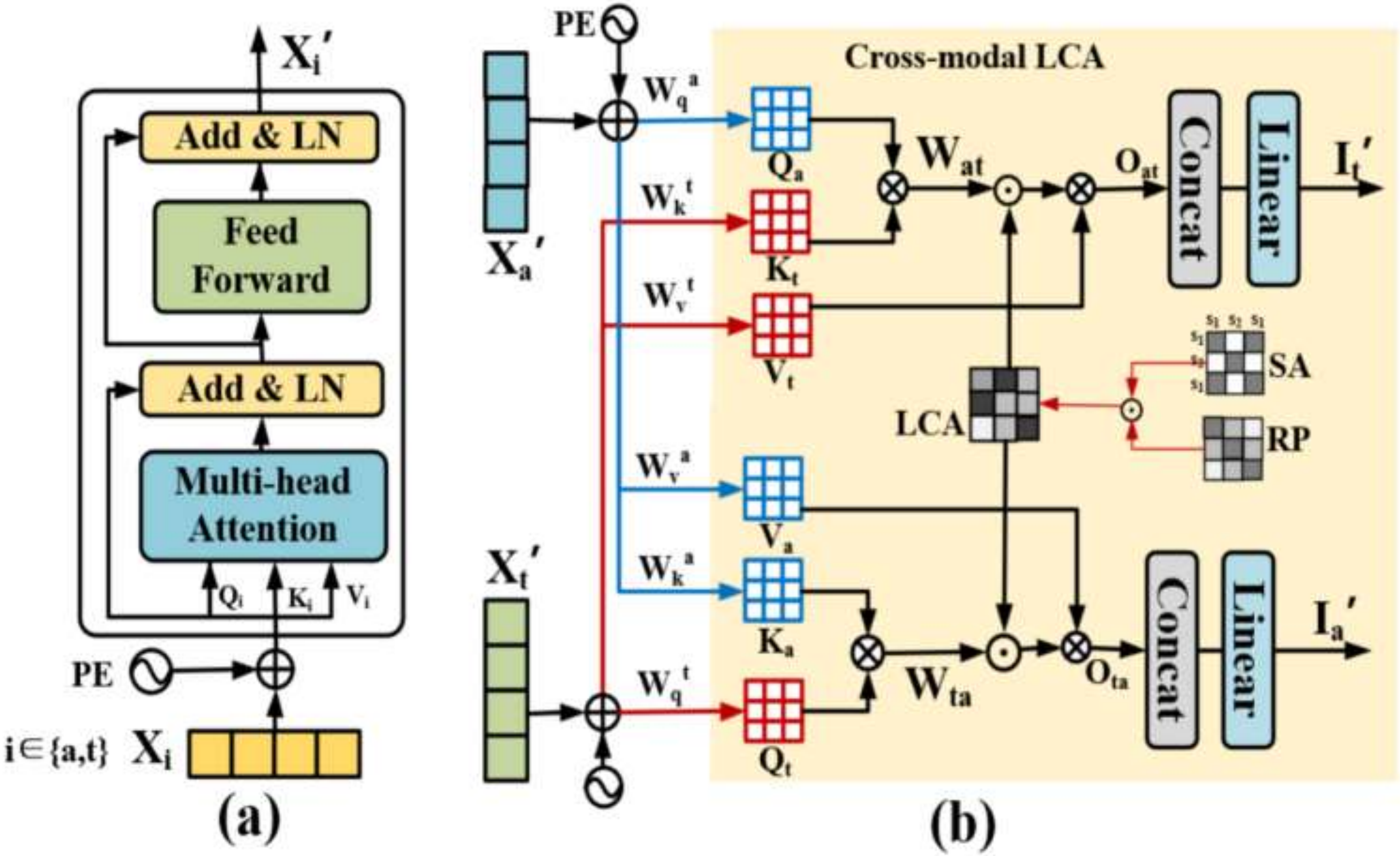
$$\mathrm{SA}_{m,n} = \begin{cases} 1 & \text{if } s_m = s_n; \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

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

$$LCA = sigmoid(RP) \times SA$$

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

Method



$$H = [H^{(1)}, H^{(2)}, ..., H^{(K)}]$$

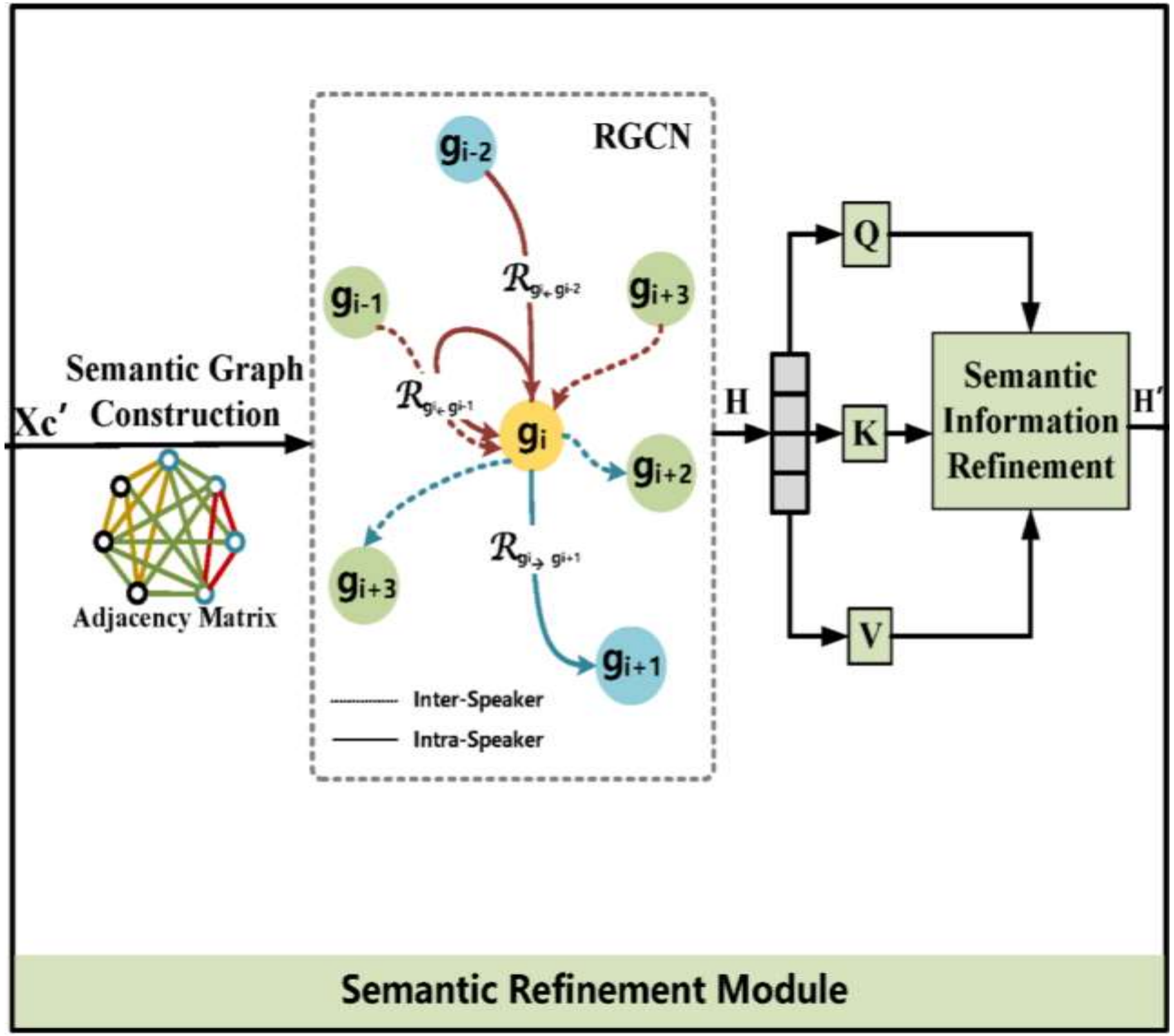
$$a_i = \mathrm{ReLU}(W^T H^{(i)} + b) \quad (7)$$

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

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

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

Method



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

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

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

Method

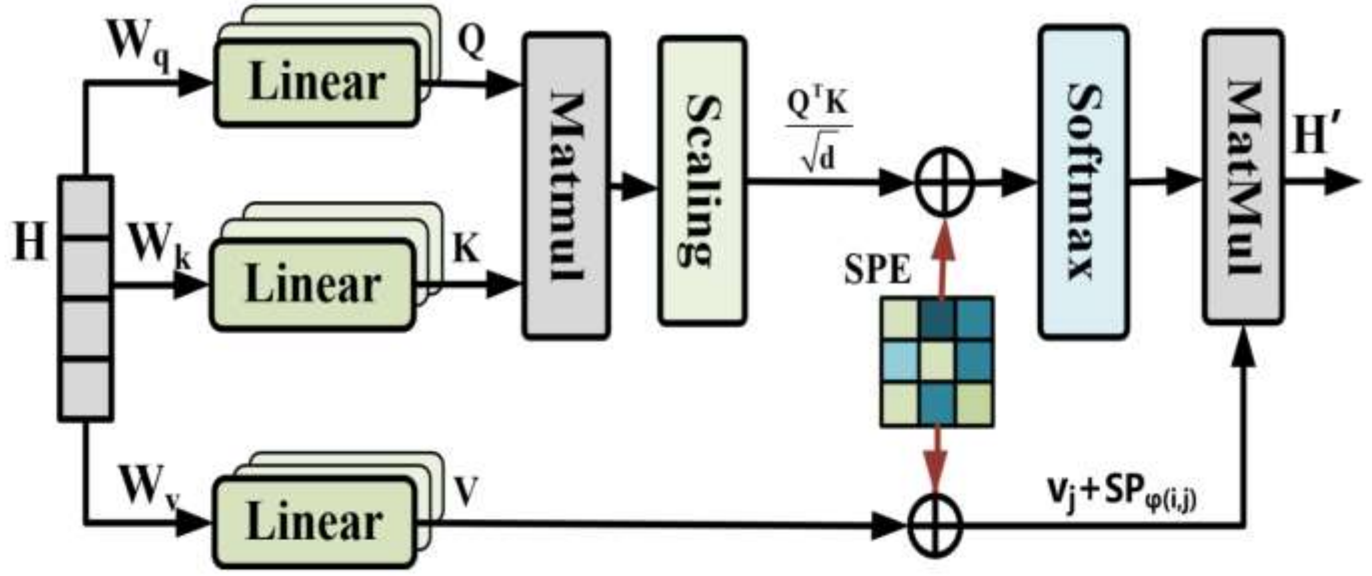


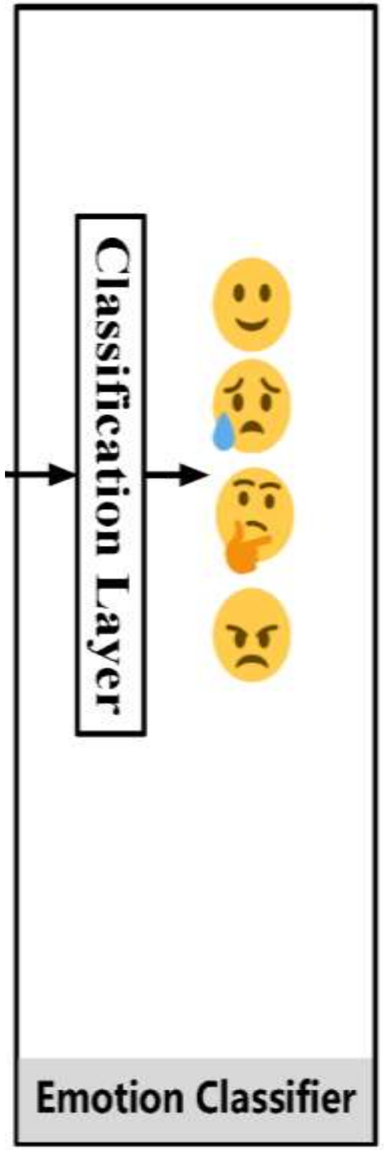
Fig. 4. Semantic Information Refinement.

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

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

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

Method



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

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

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

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

Experiments

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

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

Experiments

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**

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

Experiments

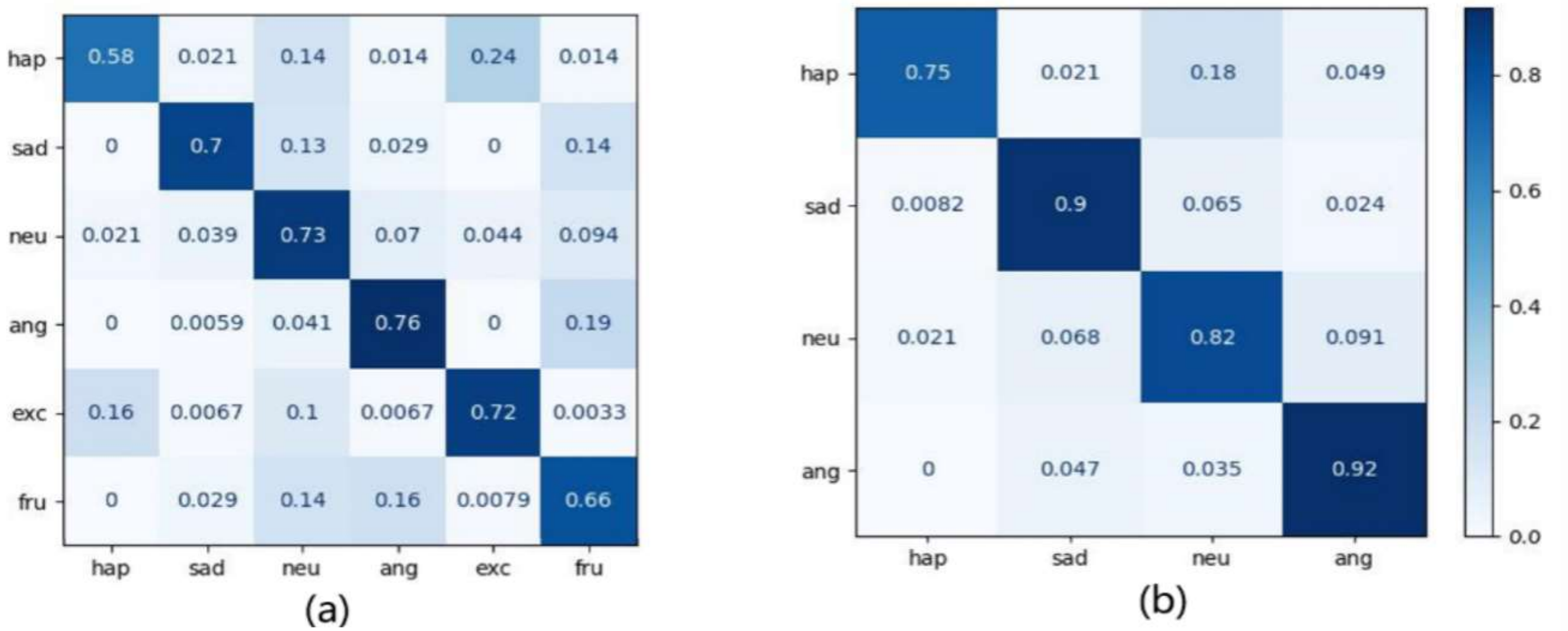


Fig. 5. Visualization the confusion matrices: (a) on the IEMOCAP (6-way); (b) on the IEMOCAP (4-way).

Experiments

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

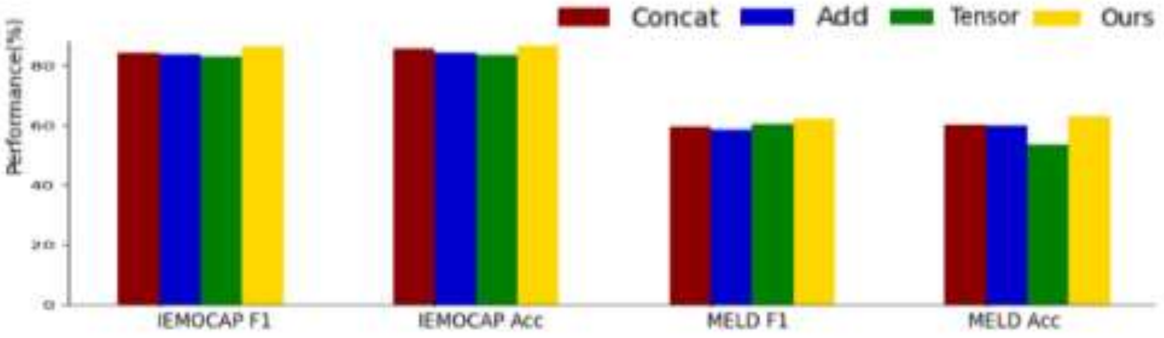


Fig. 7. Performance of different fusion strategies compared with ASB on MELD and IEMOCAP.

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**

Experiments

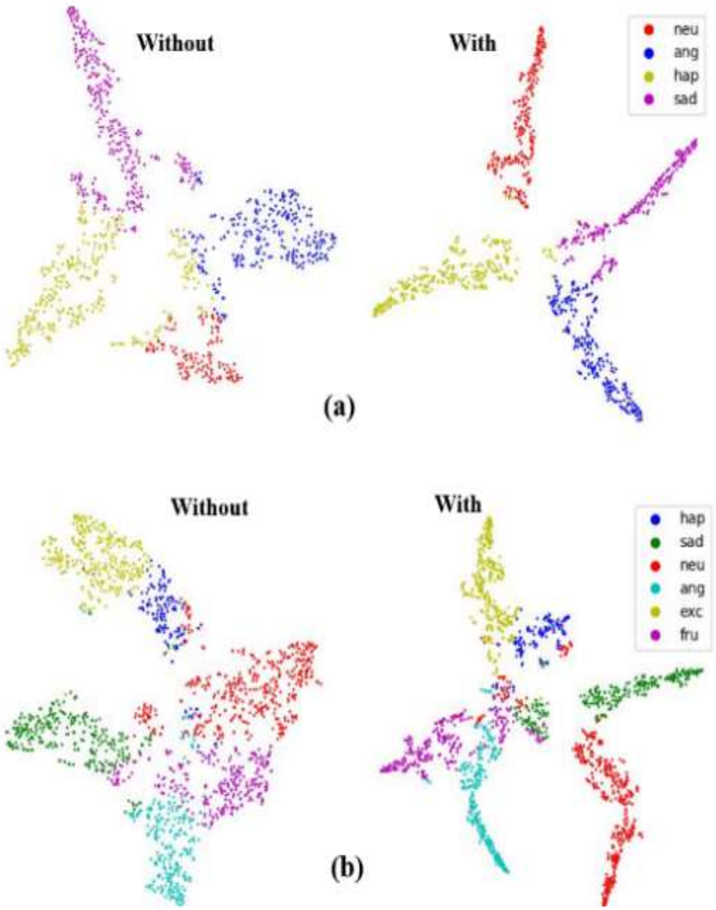


Fig. 8. T-SNE representation with and without semantic information refinement components respectively on (a) IEMOCAP (4-way) and (b) IEMOCAP (6-way).

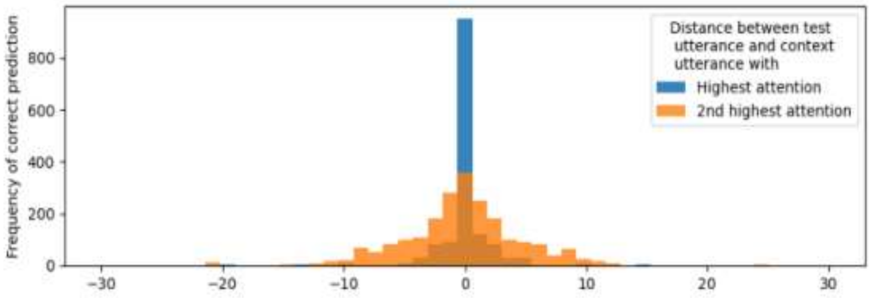


Fig. 9. Histogram of distance between the target utterance and its (2nd) highest attended utterance on MELD.

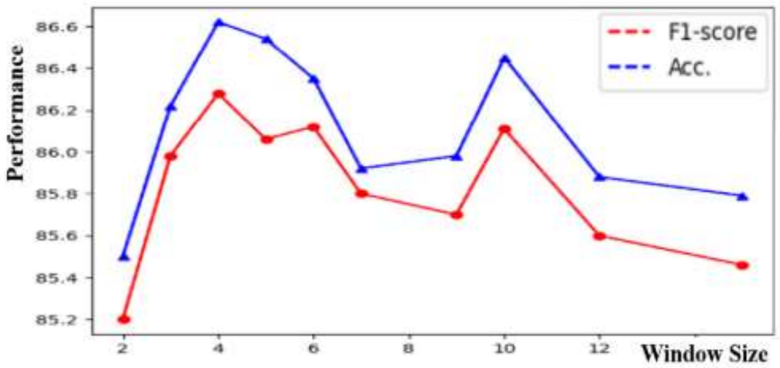


Fig. 10. Comparison for various window sizes.

Thanks!