GraphCFC: A Directed Graph based Cross-modal Feature Complementation Approach for Multimodal Conversational Emotion Recognition

Jiang Li, Xiaoping Wang, *Senior Member, IEEE*, Guoqing Lv, and Zhigang Zeng, *Fellow, IEEE*

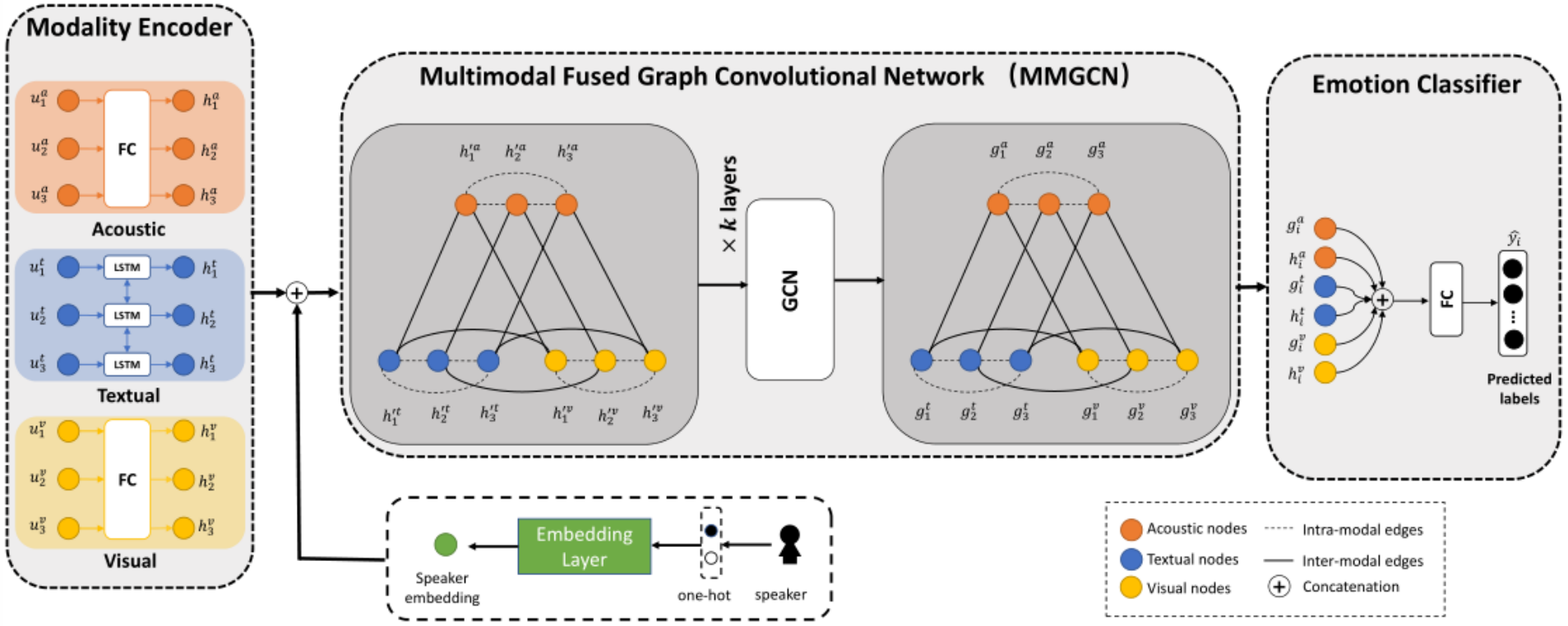
Code:None

—— IEEE Transactions on Multimedia 2023

2024. 2. 1 • ChongQing

Reported by JiaWei Cheng

Motivation



MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation(ACL2021)

This approach not only adds redundant information due to inconsistent data distribution among modalities, but also may risk losing diverse information in the conversational graph

Overview

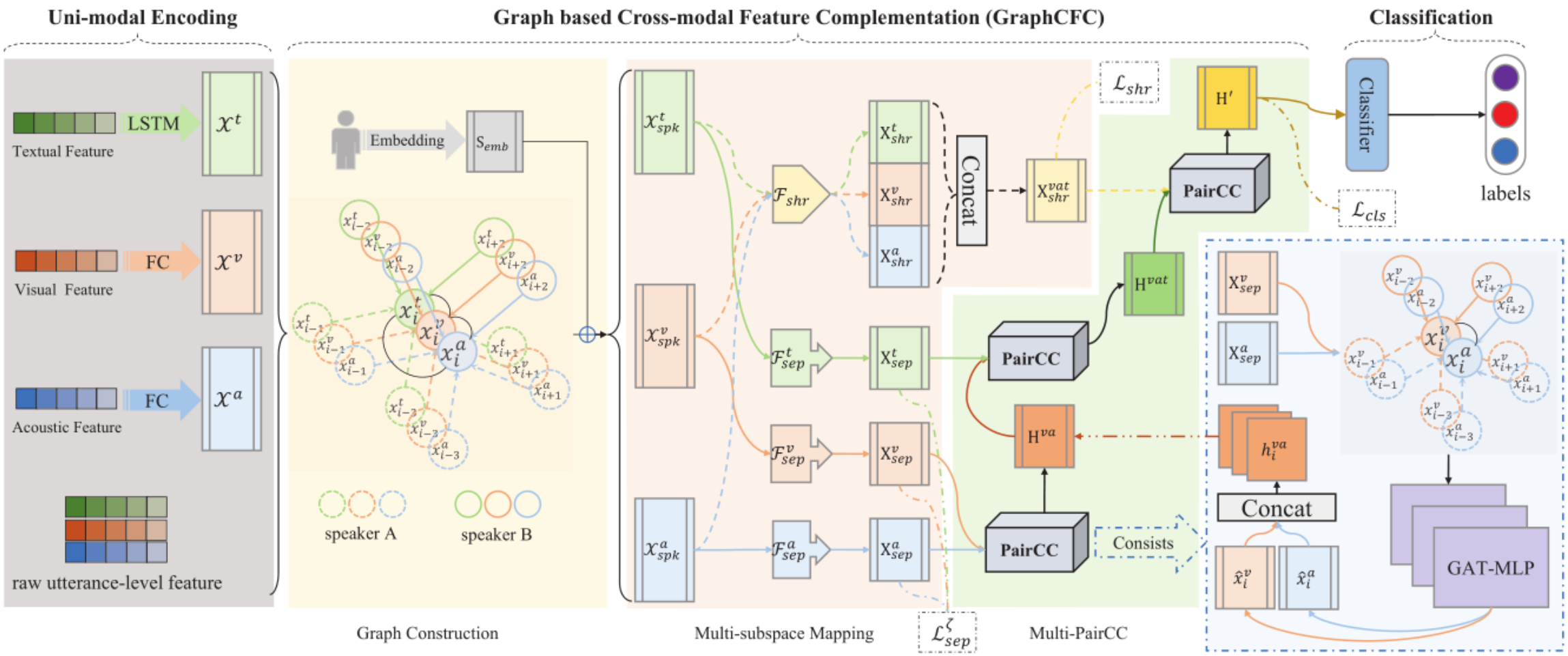
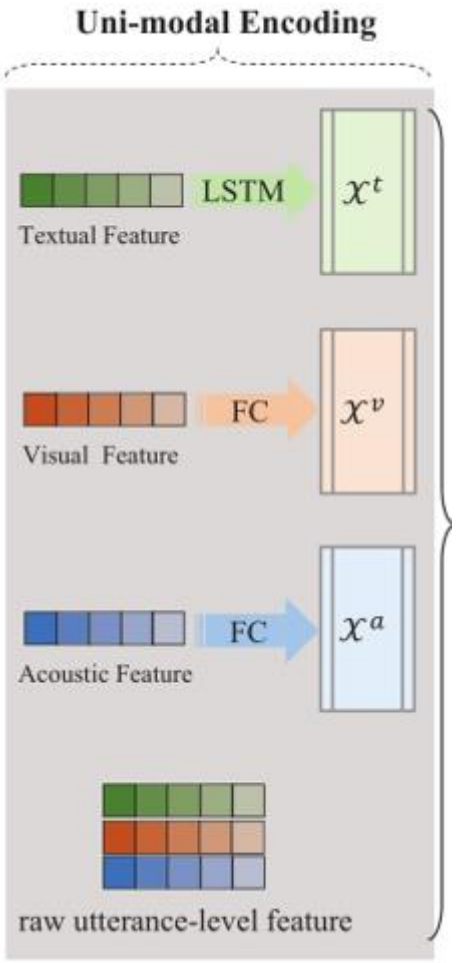


Fig. 2: The illustration of graph-based multimodal ERC, which includes uni-modal encoding, graph based cross-modal feature complementation and multimodal emotion classification.

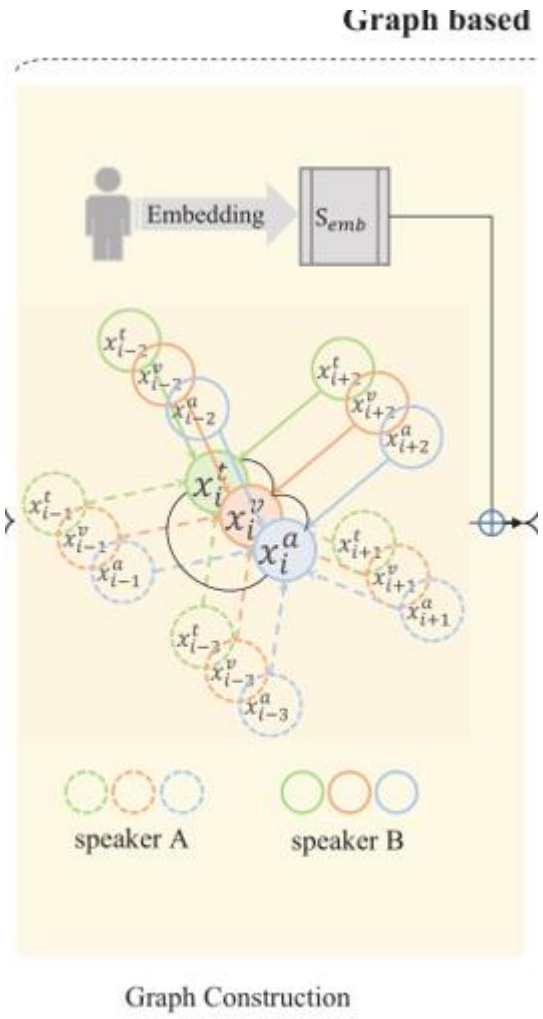
Method



$$x_i^t, x_{h,i}^t = \overleftrightarrow{\mathrm{LSTM}}(u_i^t; \Theta_{ls}^t), \tag{1}$$

$$x_i^\tau = \mathrm{FC}(u_i^\tau; \Theta_{fc}^\tau), \tau \in \{a, v\}, \tag{2}$$

Method



Graph Construction

$$\mathcal{E}_{intra} = \begin{cases} \{(u_t^{\mathrm{P}}, u_i^{\mathrm{P}}) | i-j < t < i-1\} \\ \{(u_i^{\mathrm{P}}, u_t^{\mathrm{P}}) | i+1 < t < i+k\} \end{cases}, \tag{3}$$

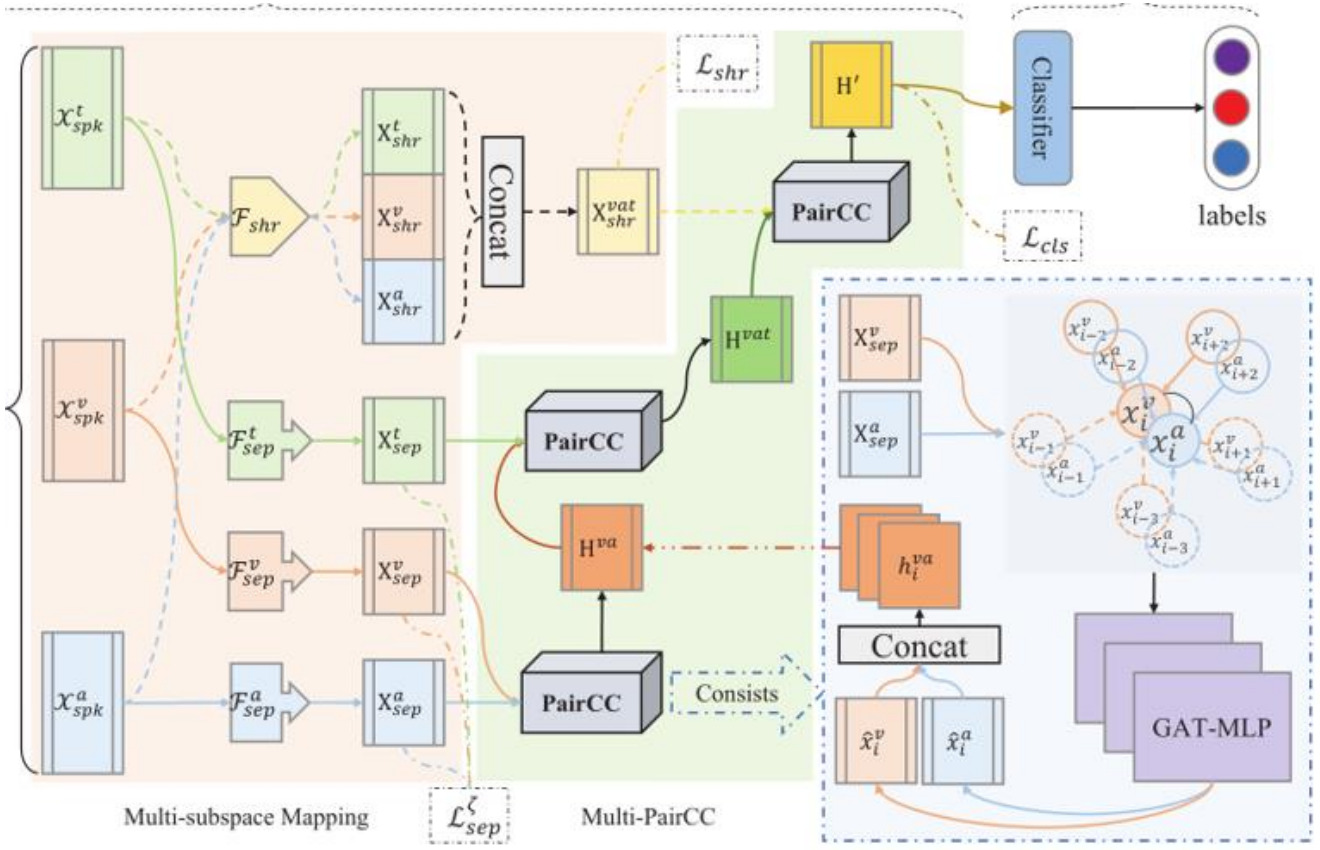
$$\mathcal{E}_{inter} = \{(u_i^{\mathrm{P}}, u_i^{\mathrm{Q}}), (u_i^{\mathrm{Q}}, u_i^{\mathrm{P}})\}, \tag{4}$$

$$\begin{aligned} \mathrm{ET}_{intra} = \{&\mathrm{et}(s_1, s_1), \mathrm{et}(s_1, s_2), \mathrm{et}(s_1, s_3), \\ &\mathrm{et}(s_2, s_2), \mathrm{et}(s_2, s_3), \mathrm{et}(s_3, s_3)\}. \end{aligned} \tag{5}$$

$$\begin{aligned} \mathrm{ET}_{inter} = \{&\mathrm{et}(mod_1, mod_2), \mathrm{et}(mod_1, mod_3), \\ &\mathrm{et}(mod_2, mod_3)\}. \end{aligned} \tag{6}$$

$$S_{emb} = \mathrm{Embedding}(\mathrm{S}, D), \tag{7}$$

$$\mathcal{X}_{spk}^{\zeta} = \mu S_{emb} + \mathcal{X}^{\zeta}, \tag{8}$$



$$
\begin{aligned}
\mathrm{X}^{\zeta}_{shr} &= \mathcal{F}_{shr}(\mathcal{X}^{\zeta}_{spk};\Theta_{shr}),\\
\mathrm{X}^{vat}_{shr} &= \mathrm{Lin}([\mathrm{X}^{v}_{shr} \parallel \mathrm{X}^{a}_{shr} \parallel \mathrm{X}^{t}_{shr}];\Theta'_{shr}),\\
\mathrm{X}^{\zeta}_{sep} &= \mathcal{F}_{sep}(\mathcal{X}^{\zeta}_{spk};\Theta^{\zeta}_{sep}),
\end{aligned}
\tag{9}
$$

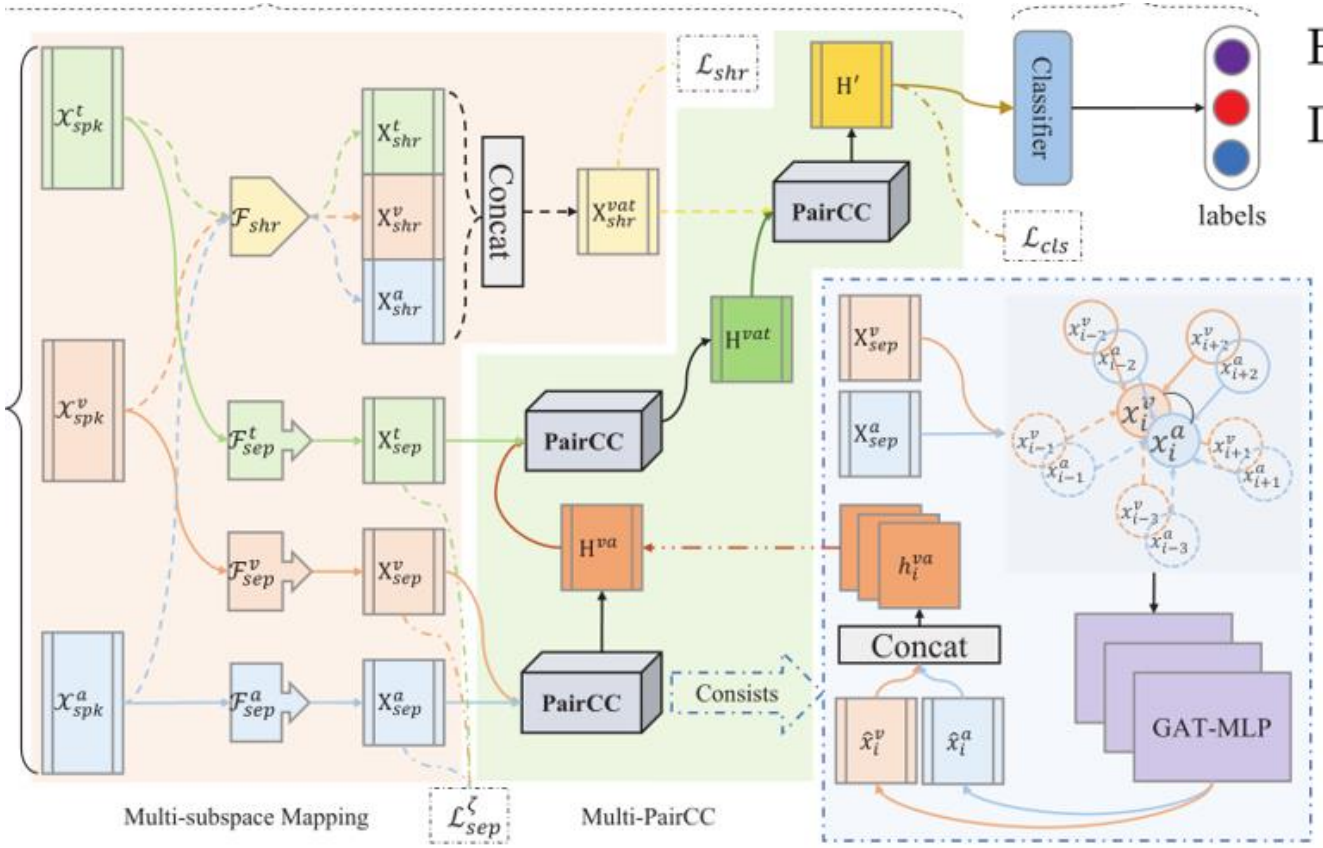
$$
\begin{aligned}
&\mathcal{F}(\mathcal{X};\Theta) =\\
&\mathrm{Norm}(\mathrm{Drop}(\mathrm{Lin}(\mathrm{Drop}(\sigma(\mathrm{Lin}(\mathcal{X};\Theta_0)));\Theta_1))),
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
v'_i &= \mathrm{ReLU}(\mathrm{W}'_0\mathrm{x}^{vat}_{shr,i} + \mathrm{b}'_0),\\
p'_i &= \mathrm{Softmax}(\mathrm{W}'_1 v^s_i + \mathrm{b}'_1),
\end{aligned}
\tag{11}
$$

$$
\mathcal{L}_{shr} = -\frac{1}{\sum_{k=0}^{N-1} n(k)} \sum_{i=0}^{N-1} \sum_{j=0}^{n(i)-1} y_{ij} \log p'_{ij} + \lambda|\Theta'_{re}|, \tag{12}
$$

$$
\begin{aligned}
v^{\zeta}_i &= \mathrm{ReLU}(\mathrm{W}^{\zeta}_0\mathrm{x}^{\zeta}_{sep,i} + \mathrm{b}^{\zeta}_0),\\
p^{\zeta}_i &= \mathrm{Softmax}(\mathrm{W}^{\zeta}_1 v^{\zeta}_i + \mathrm{b}^{\zeta}_1),
\end{aligned}
\tag{13}
$$

$$
\mathcal{L}^{\zeta}_{sep} = -\frac{1}{\sum_{k=0}^{N-1} n(k)} \sum_{i=0}^{N-1} \sum_{j=0}^{n(i)-1} y_{ij} \log p^{\zeta}_{ij} + \lambda|\Theta^{\zeta}_{re}|, \tag{14}
$$



$$\begin{aligned}\mathrm{X}_{\mathrm{gat}} &= \mathrm{Norm}(\mathrm{MultiGAT}(\mathcal{E}, \mathrm{X}_{in}; \Theta_{\mathrm{gat}}) + \mathrm{X}_{in}),\\ \mathrm{X}_{out} &= \mathrm{Norm}(\mathrm{FeedForward}(\mathrm{X}_{\mathrm{gat}}; \Theta_{\mathrm{fed}}) + \mathrm{X}_{\mathrm{gat}}),\end{aligned} \tag{15}$$

$$\begin{aligned}&\mathrm{FeedForward}(\mathrm{X}_{\mathrm{gat}}, \Theta_{\mathrm{fed}}) =\\ &\mathrm{Drop}(\mathrm{Lin}(\mathrm{Drop}(\sigma(\mathrm{Lin}(\mathrm{X}_{\mathrm{gat}}; \Theta_0))); \Theta_1)),\end{aligned} \tag{16}$$

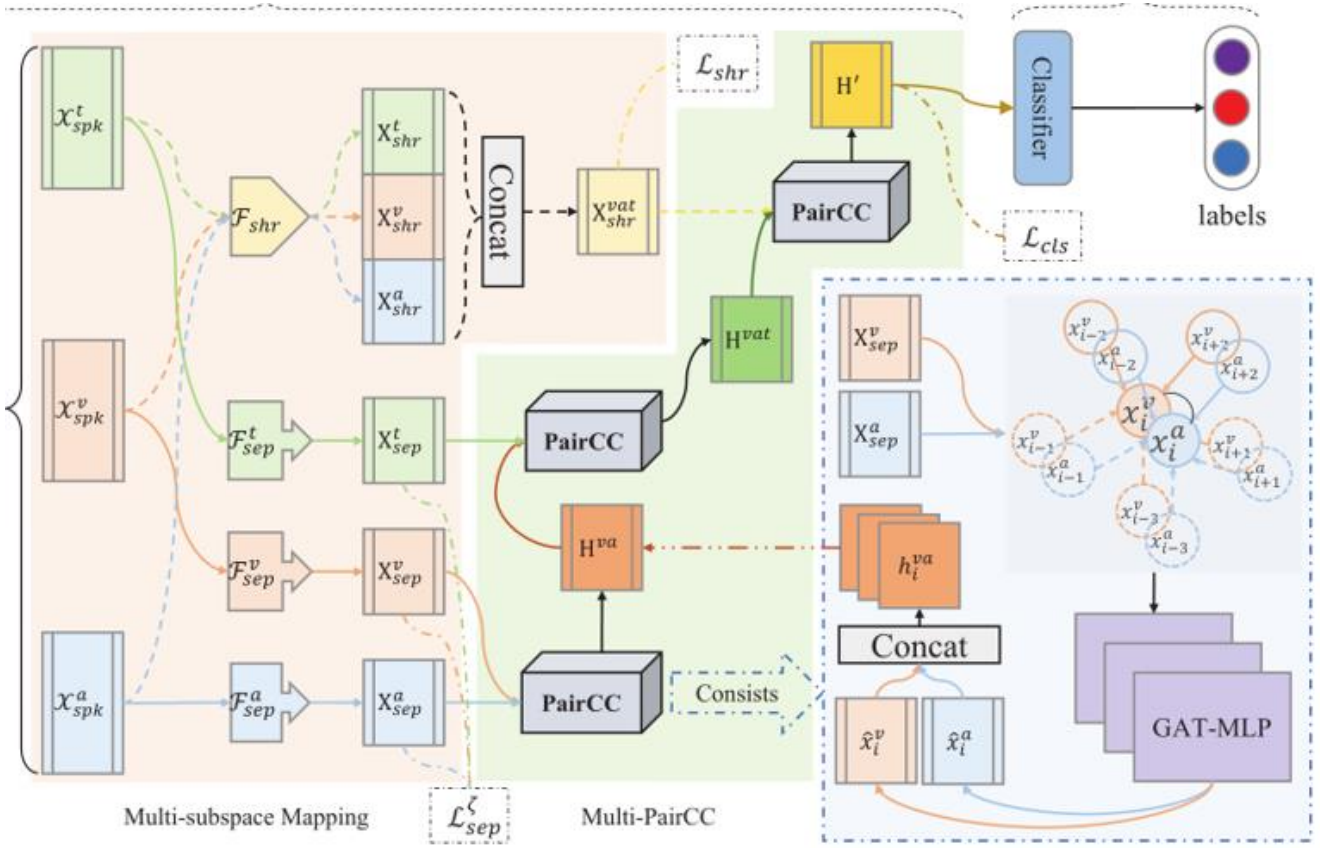
$$\begin{aligned}&\mathrm{MultiGAT}(\mathcal{E}, \mathrm{X}_{in}; \Theta_h) = \Theta_h[\mathrm{head}_1 \,\|\, \ldots \,\|\, \mathrm{head}_h],\\ &\textit{where } \mathrm{head}_i = \mathrm{SingleGAT}(\mathcal{E}, \mathrm{X}_{in}; \Theta_i),\end{aligned} \tag{17}$$

$$\begin{aligned}\mathrm{X}_{\mathrm{gat}} &= \mathrm{MultiGAT}(\mathcal{E}, \mathrm{Norm}(\mathrm{X}_{in}); \Theta_{\mathrm{gat}}) + \mathrm{X}_{in},\\ \mathrm{X}_{out} &= \mathrm{FeedForward}(\mathrm{Norm}(\mathrm{X}_{\mathrm{gat}}); \Theta_{\mathrm{fed}}) + \mathrm{X}_{\mathrm{gat}}.\end{aligned} \tag{18}$$

$$\begin{aligned}\mathrm{H}^{va} &= \mathrm{PairCC}(\mathrm{X}^{v}_{sep}, \mathrm{X}^{a}_{sep}; \Theta^{va}_{sep}),\\ \mathrm{H}^{vat} &= \mathrm{PairCC}(\mathrm{X}^{t}_{sep}, \mathrm{H}^{va}; \Theta^{vat}_{sep}),\\ \mathrm{H}' &= \mathrm{PairCC}(\mathrm{X}^{vat}_{shr}, \mathrm{H}^{vat}; \Theta'),\end{aligned} \tag{19}$$

$$\begin{aligned}\mathrm{x}_{agg,i} &= \mathrm{AGG}(\{\mathrm{x}_j | w_j \in \mathcal{N}(w_i)\}; \Theta_{agg}),\\ \mathrm{x}_{com,i} &= \mathrm{COM}(\mathrm{x}_i, \mathrm{x}_{agg,i}; \Theta_{com}),\end{aligned} \tag{20}$$

$$\mathrm{x}_{agg,i} = \sum_{w_j \in \mathcal{N}(w_i)} \alpha_{ij} \mathrm{W}_{agg} \mathrm{x}_j, \tag{21}$$



$$\alpha_{ij} = \frac{\exp\left(\mathrm{a}^{\top}\sigma\left(\Theta_{\mathrm{att}}[\mathrm{x}_i \,\|\, \mathrm{x}_j]\right)\right)}{\sum_{w_k \in \mathcal{N}(w_i)} \exp\left(\mathrm{a}^{\top}\sigma\left(\Theta_{\mathrm{att}}[\mathrm{x}_i \,\|\, \mathrm{x}_k]\right)\right)}, \tag{22}$$

$$\begin{aligned} &ET_{emb} = \mathrm{Embedding}(\mathrm{ET}, DM), \\ &where\ DM = M \times (D^2 + D + M - 1)/2, \end{aligned} \tag{23}$$

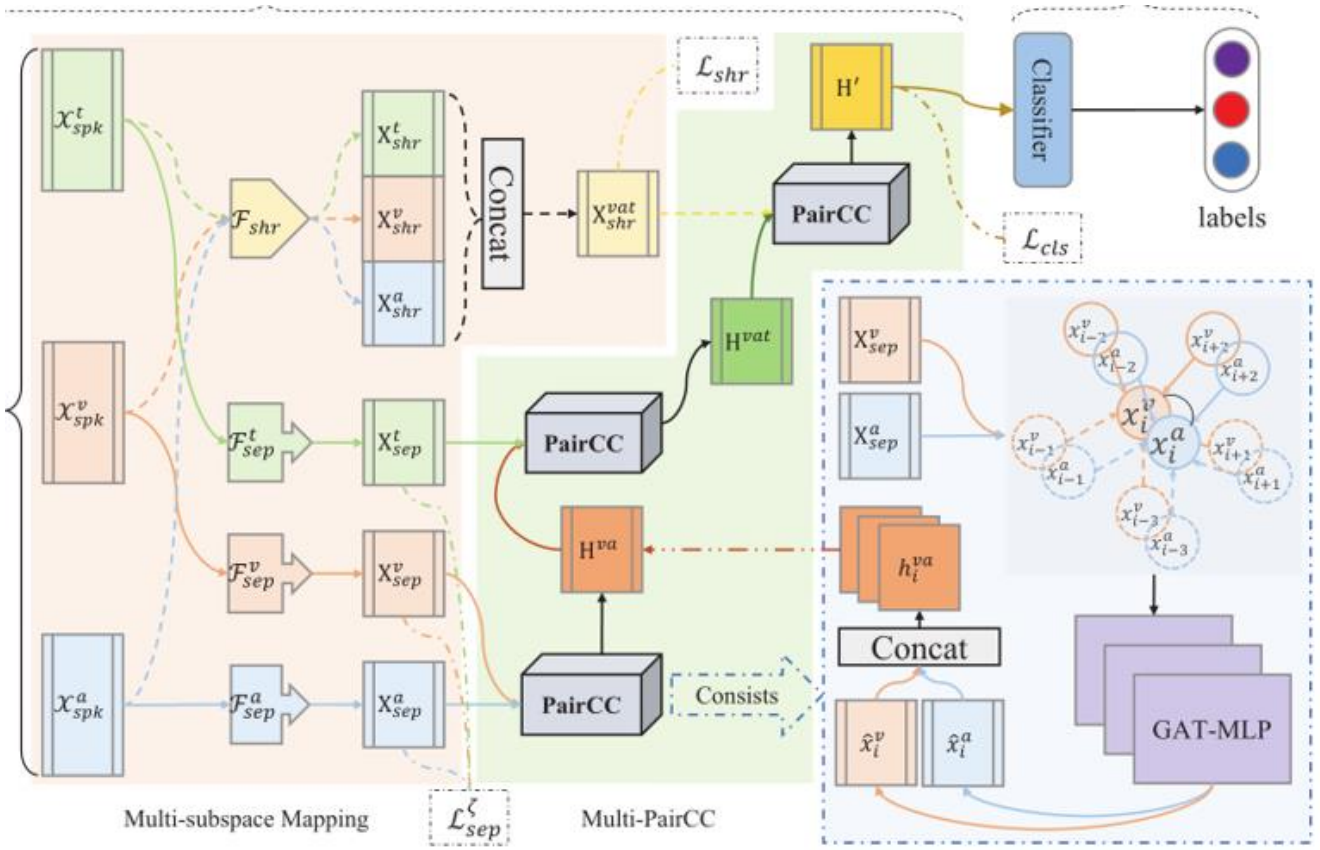
$$\alpha'_{ij} = \frac{\exp\left(\mathrm{a}^{\top}\sigma\left(\Theta_{\mathrm{att}}[\mathrm{x}_i \,\|\, \mathrm{x}_j \,\|\, \mathrm{et}_{ij}]\right)\right)}{\sum_{w_k \in \mathcal{N}(w_i)} \exp\left(\mathrm{a}^{\top}\sigma\left(\Theta_{\mathrm{att}}[\mathrm{x}_i \,\|\, \mathrm{x}_k \,\|\, \mathrm{et}_{ik}]\right)\right)}, \tag{24}$$

$$\mathrm{x}^1_{com,i} = \mathrm{GRU}(\mathrm{x}_i, \mathrm{x}_{agg,i}; \Theta^1_{com}), \tag{25}$$

$$\mathrm{x}^2_{com,i} = \mathrm{GRU}(\mathrm{x}_{agg,i}, \mathrm{x}_i; \Theta^2_{com}), \tag{26}$$

$$\mathrm{x}_{com,i} = \mathrm{x}^1_{com,i} + \mathrm{x}^2_{com,i}, \tag{27}$$

$$\mathrm{x}_{\mathrm{gat},i} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{x}^k_{com,i}, \tag{28}$$



$$
\begin{aligned}
v_i &= \mathrm{ReLU}(\mathrm{W}_0 h_i + \mathrm{b}_0), \\
p_i &= \mathrm{Softmax}(\mathrm{W}_1 v_i + \mathrm{b}_1), \\
\hat{y}_i &= \underset{\mathrm{k}}{\mathrm{argmax}}(p_i[\mathrm{k}]),
\end{aligned} \tag{29}
$$

$$
\mathcal{L}_{cls} = -\frac{1}{\sum_{k=0}^{N-1} n(k)} \sum_{i=0}^{N-1} \sum_{j=0}^{n(i)-1} y_{ij} \log p_{ij} + \lambda |\Theta_{re}|, \tag{30}
$$

$$
\mathcal{L} = \mathcal{L}_{cls} + \beta \mathcal{L}_{shr} + \gamma^a \mathcal{L}^a_{sep} + \gamma^v \mathcal{L}^v_{sep} + \gamma^t \mathcal{L}^t_{sep}, \tag{31}
$$

Experiments

Model	IEMOCAP								MELD						
	Happy	*Sad*	*Neutral*	*Angry*	*Excited*	*Frustrated*	Accuracy	wa-F1	*Neutral*	*Surprise*	*Sadness*	*Joy*	*Anger*	Accuracy	wa-F1
bc-LSTM	32.63	70.34	51.14	63.44	67.91	61.06	59.58	59.10	75.66	48.47	22.06	52.10	44.39	59.62	56.80
CMN	30.38	62.41	52.39	59.83	60.25	60.69	56.56	56.13	-	-	-	-	-	-	-
ICON	29.91	64.57	57.38	63.04	63.42	60.81	59.09	58.54	-	-	-	-	-	-	-
DialogueRNN	33.18	78.80	59.21	65.28	71.86	58.91	63.40	62.75	76.79	47.69	20.41	50.92	45.52	60.31	57.66
DialogueCRN	**51.59**	74.54	62.38	67.25	73.96	59.97	65.31	65.34	76.13	46.55	11.43	49.47	44.92	59.66	56.76
DialogueGCN	47.10	80.88	58.71	66.08	70.97	61.21	65.54	65.04	75.97	46.05	19.60	51.20	40.83	58.62	56.36
MMGCN	45.45	77.53	61.99	66.67	72.04	**64.12**	65.56	65.71	75.16	48.45	25.71	**54.41**	45.45	59.31	57.82
GraphCFC	43.08	**84.99**	**64.70**	**71.35**	**78.86**	63.70	**69.13**	**68.91**	**76.98**	**49.36**	**26.89**	51.88	**47.59**	**61.42**	**58.86**

Experiments

Modality Setting	IEMOCAP		MELD	
	Accuracy	wa-F1	Accuracy	wa-F1
A	54.16	53.85	47.55	41.62
V	31.61	27.67	47.59	33.26
T	59.95	60.09	60.77	56.81
A + V	54.17	53.89	47.61	41.67
A + T	64.20	64.74	59.96	57.46
V + T	63.15	62.96	59.46	57.29
A + V + T	**69.13**	**68.91**	**61.42**	**58.86**

Experiments

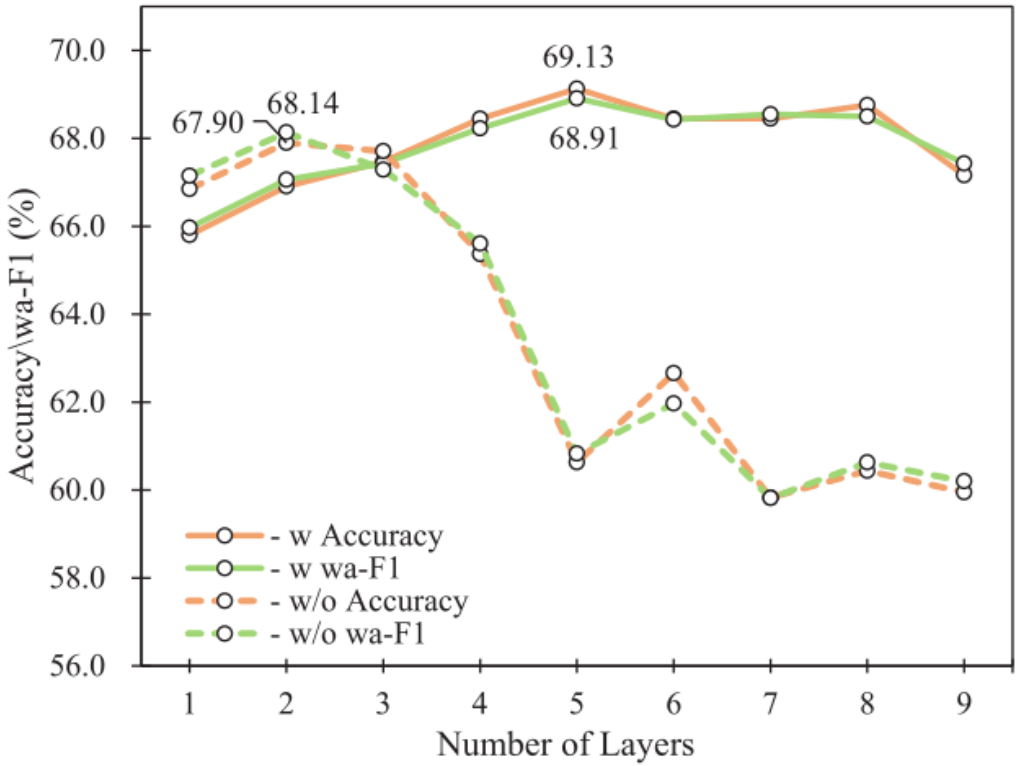


Fig. 4: The effects of the number of GAT-MLP layer and skip connection on our model. The figure shows the results on the IEMOCAP dataset. w (w/o) indicates the use (non-use) of skip connection.

Experiments

TABLE V: The effect of the multi-subspace loss functions. $\mathcal{L}_{shr}$ and $\mathcal{L}_{sep}^{\zeta}$ ($\zeta \in \{a, v, t\}$) denote shared and separate subspace losses, respectively.

$\mathcal{L}_{shr}$	$\mathcal{L}_{sep}^{\zeta}$	IEMOCAP		MELD	
		Accuracy	wa-F1	Accuracy	wa-F1
- w/o	- w	68.70	68.35	61.00	58.39
- w	- w/o	67.53	67.56	60.27	57.99
- w/o	- w/o	68.70	68.36	60.38	58.09
- w	- w	**69.13**	**68.91**	**61.42**	**58.86**

Experiments

TABLE VI: The influence of speakers and edge types on our GraphCFC model. S_{emb} and E_{emb} indicate the embeddings of multi-speaker and edge types, respectively.

S_{emb}	E_{emb}	IEMOCAP		MELD	
		Accuracy	wa-F1	Accuracy	wa-F1
- w/o	- w	68.02	68.04	60.69	58.35
- w	- w/o	65.26	65.91	60.46	57.91
- w	- w	**69.13**	**68.91**	**61.42**	**58.86**

Experiments

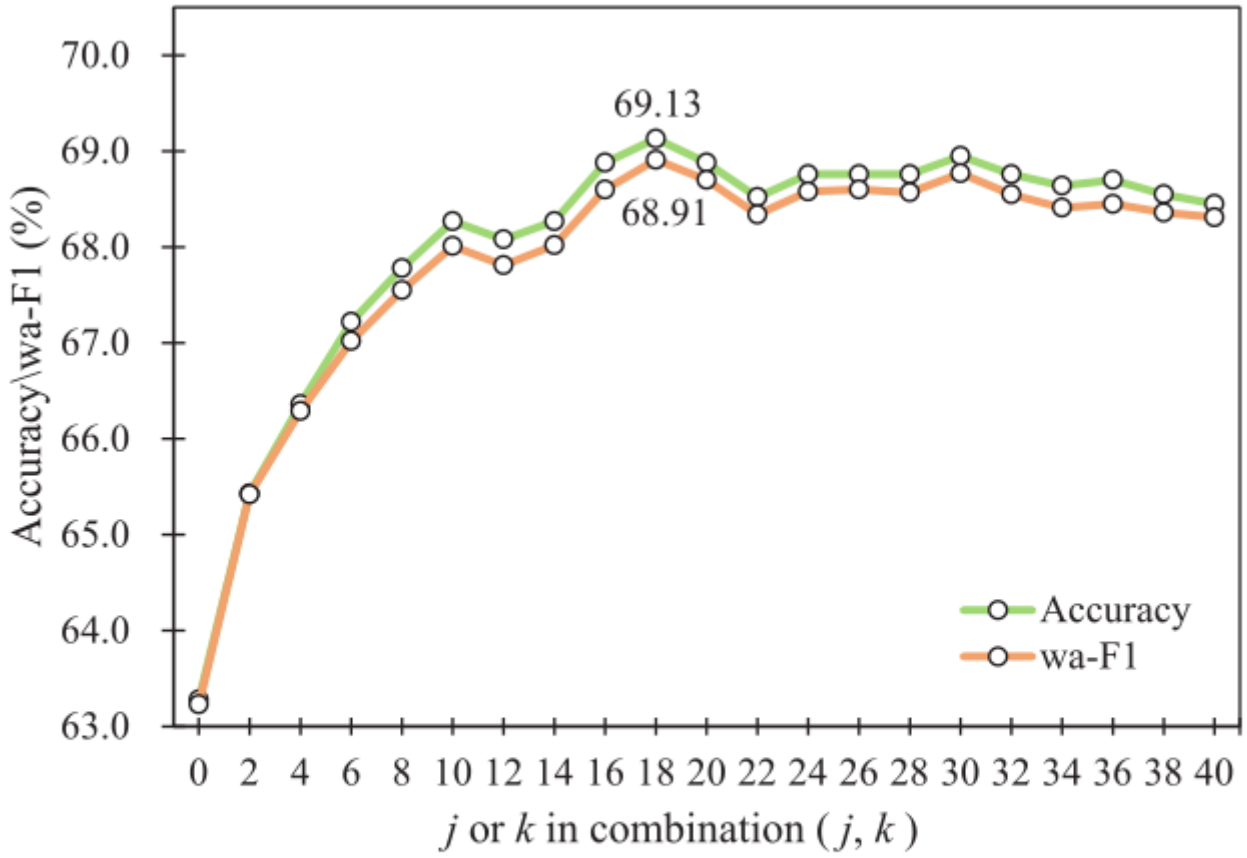


Fig. 5: The effects of j nodes in the past and k nodes in the future on the proposed GraphCFC model. The figure shows the results on the IEMOCAP dataset.

Experiments

TABLE VII: The statistics of the merged emotion labels.

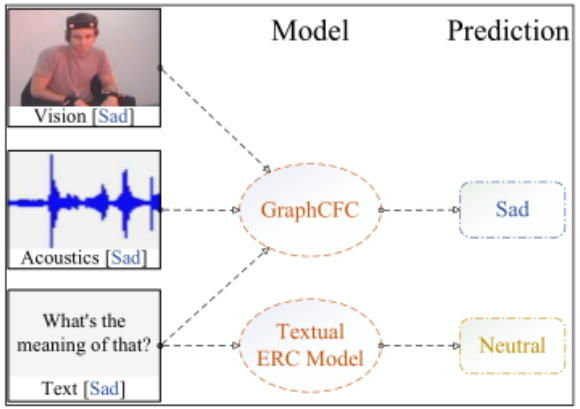
New Label	IEMOCAP	MELD
Positive	*Happy, Excited*	*Joy*
Negative	*Sad, Angry, Frustrated*	*Surprise, Fear, Sadness, Disgust, Anger*
Neutral	*Neutral*	*Neutral*

Experiments

TABLE VIII: The overall performance after converting the dataset into three-emotion labels under the multimodal setting.

Model	IEMOCAP					MELD				
	Positive	*Neutral*	*Negative*	Accuracy	wa-F1	*Positive*	*Neutral*	*Negative*	Accuracy	wa-F1
bc-LSTM	**90.58**	55.63	84.04	79.54	79.10	36.97	75.12	61.46	65.13	64.26
DialogueRNN	88.36	57.99	83.81	78.87	78.94	40.29	74.95	62.10	65.52	64.93
DialogueCRN	79.39	61.51	83.09	75.66	76.97	40.80	74.40	62.87	65.98	65.32
DialogueGCN	84.22	56.88	83.66	77.57	77.48	32.92	**75.64**	63.96	66.67	64.80
MMGCN	85.20	**64.21**	83.73	79.36	79.95	43.32	75.5	65.57	67.93	66.92
GraphCFC	88.48	62.03	**84.35**	**79.91**	**80.20**	**50.66**	75.12	**66.26**	**68.54**	**68.12**

Experiments



(a)

Turn	Speaker	Utterance[Emotion Label]	Prediction of [Baseline][GraphCFC]
17	A	Do you want my jacket?[*Neutral*]	[*Neutral*] [*Neutral*]
18	A	Right.[*Neutral*]	[*Neutral*] [*Neutral*]
19	A	Yeah, that's my point too.[*Neutral*]	[*Neutral*] [*Neutral*]
20	A	No. No, no, no, that's not my point. My point is there's no point in coming down here if we just turn around and leave.[*Excited*]	[*Neutral*] [*Excited*]

(b)

Turn	Speaker	Utterance[Emotion Label]	Prediction of [Baseline][GraphCFC]
103	A	I mean it's not really free because you have to pay for it, but yeah.[*Neutral*]	[*Neutral*] [*Neutral*]
104	A	Well do you think you're going to live on campus or off campus your first year?[*Neutral*]	[*Neutral*] [*Neutral*]
105	A	I hate to break this to you but you meet people by being cool, so-[*Neutral*]	[*Neutral*] [*Neutral*]
106	B	Oh, oh, I'm sorry, my bad. I'll have to, like, sign up for summer lessons from you before I go. You can teach me how to be cool before I go to college.[*Excited*]	[*Neutral*] [*Excited*]

(c)

Fig. 6: The cases of ERC on the IEMOCAP. (a) An example shows that multi-modality can be used to compensate for the shortcoming of single-textual modality. (b) Emotional-shift in one-speaker scenario. (c) Emotional-shift in two-speaker scenario.

Thanks!