Conversation Understanding using Relational Temporal Graph Neural Networks with Auxiliary Cross-Modality Interaction

Cam-Van Thi Nguyen[1] , **Anh-Tuan Mai**[1,2], **The-Son Le**[1]
Hai-Dang Kieu[1], **Duc-Trong Le**[1]
[1]VNU University of Engineering and Technology, Hanoi, Vietnam
[2]FPT Software AI Center
{vanntc, 20020269, 21020089, dangkh_uet, trongld}@vnu.edu.vn

2023.11.23 • ChongQing

—— EMNLP 2023

Reported by Renhui Luo

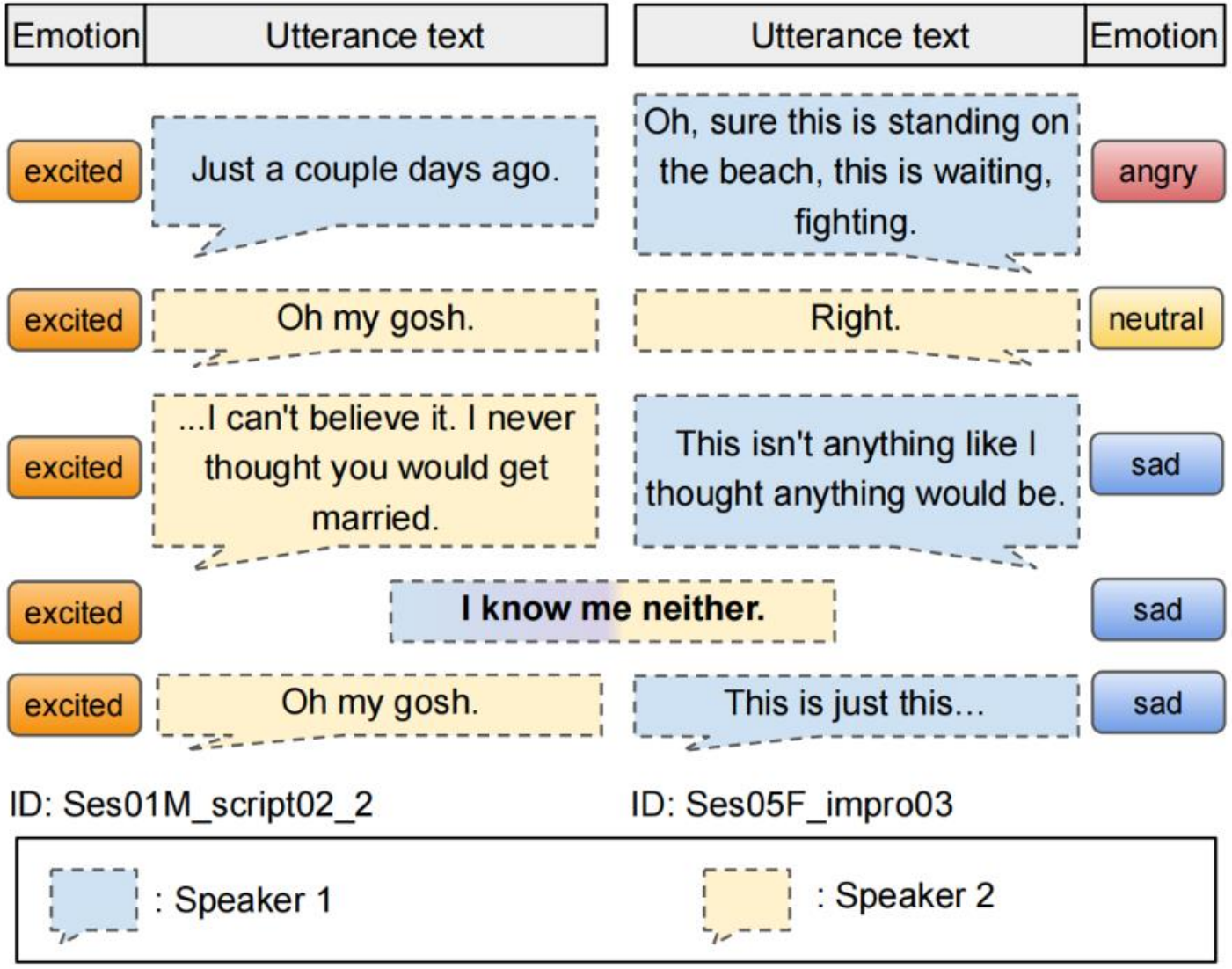
1.Introduction

2.Overview

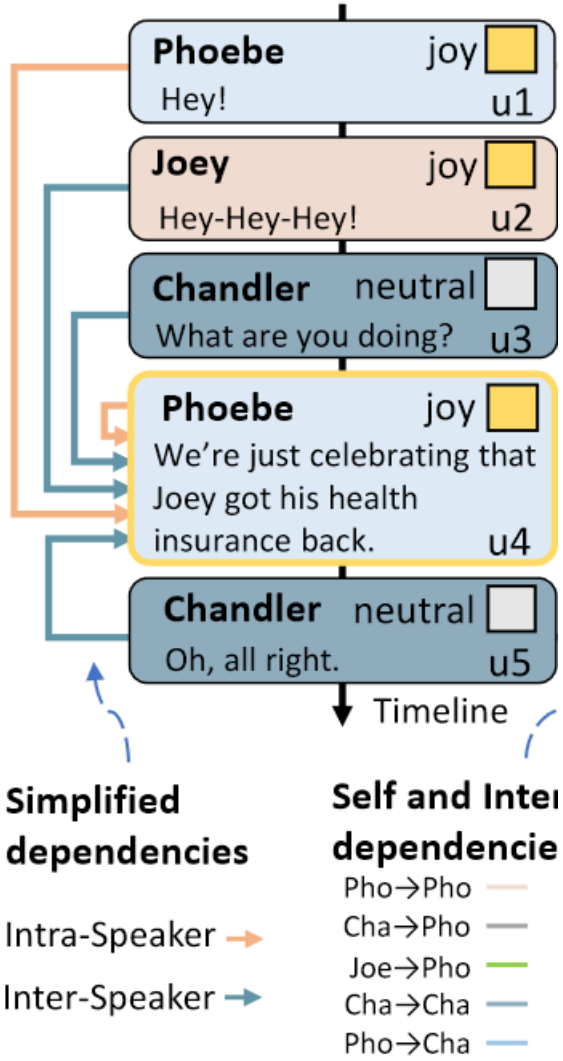
3.Methods

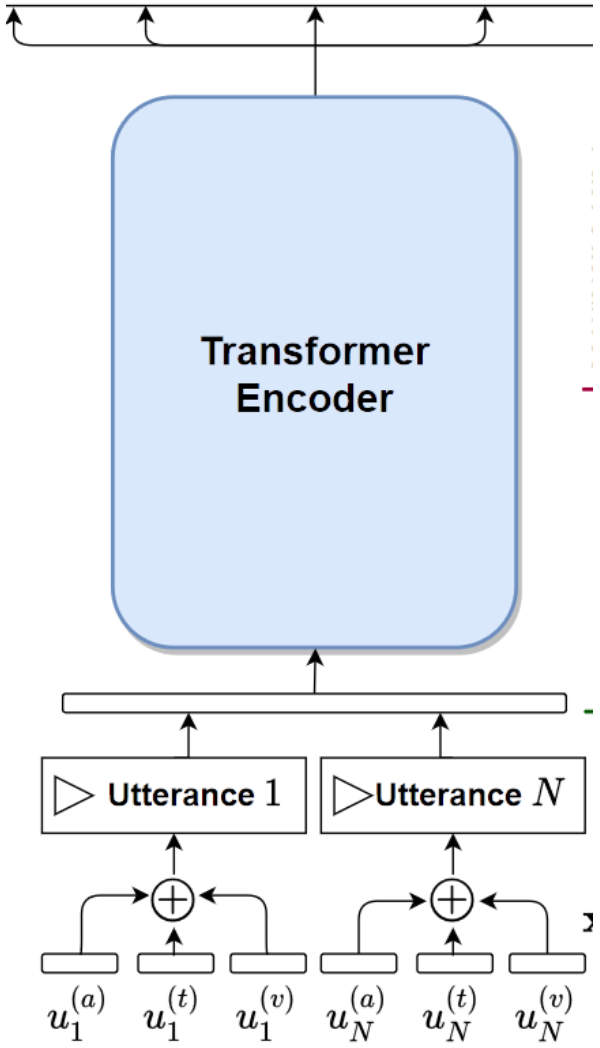
4.Experiments

Introduction

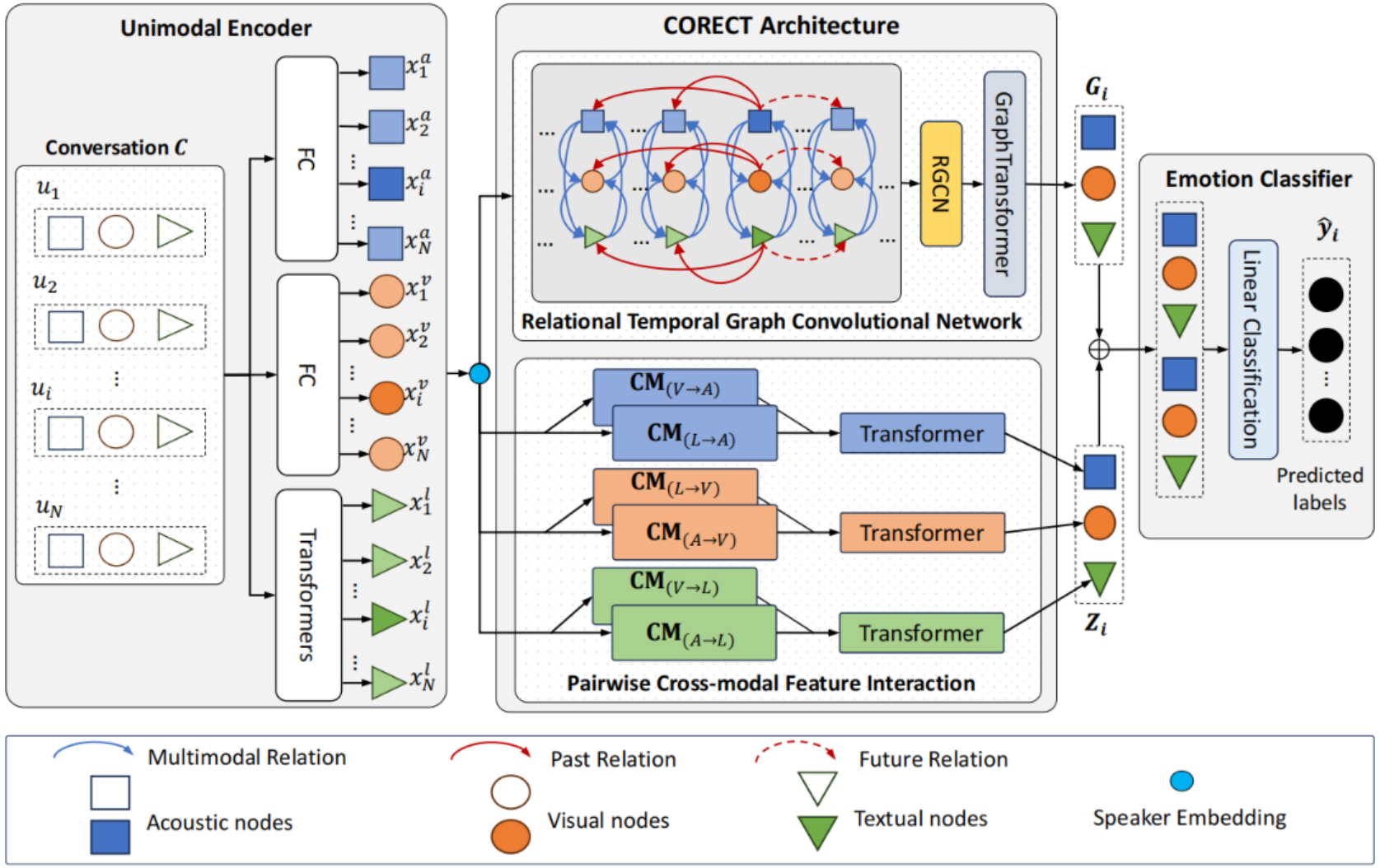


Introduction

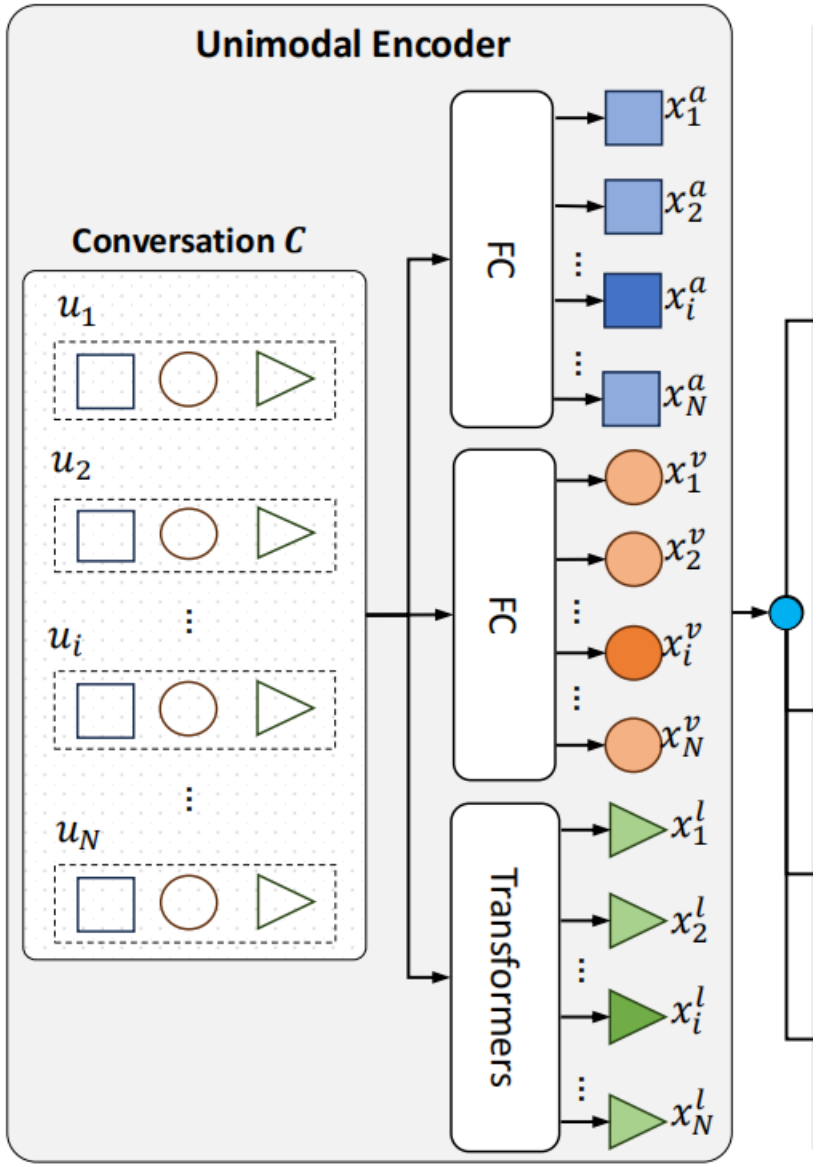




Overview


Unimodal Encoder
Conversation C
u_1
u_2
u_i
u_N
FC
FC
Transformers
x_1^a x_2^a x_i^a x_N^a
x_1^v x_2^v x_i^v x_N^v
x_1^l x_2^l x_i^l x_N^l
CORECT Architecture
RGCN
GraphTransformer
Relational Temporal Graph Convolutional Network
$CM_{(V\to A)}$
$CM_{(L\to A)}$
$CM_{(L\to V)}$
$CM_{(A\to V)}$
$CM_{(V\to L)}$
$CM_{(A\to L)}$
Transformer
Transformer
Transformer
Pairwise Cross-modal Feature Interaction
G_i
Z_i
Emotion Classifier
Linear Classification
$\hat{y}_i$
Predicted labels
Multimodal Relation
Acoustic nodes
Past Relation
Visual nodes
Future Relation
Textual nodes
Speaker Embedding

Method



audio (a), visual (v), and textual (l)

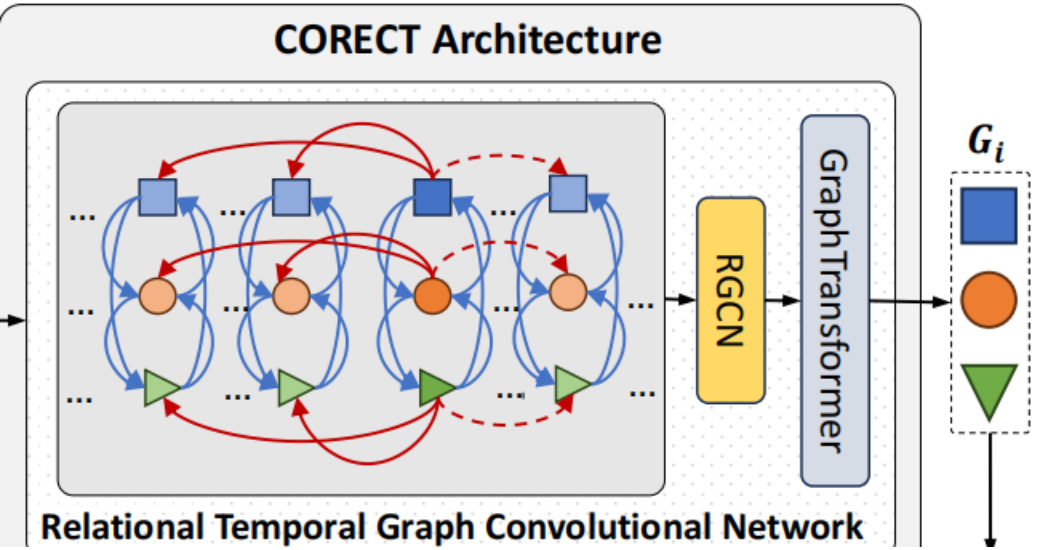
$$x_i^l = \textbf{Transformer}(u_i^l, \mathbf{W}_{trans}^l) \quad (1)$$

$$x_i^\tau = \textbf{FC}(u_i^\tau; \mathbf{W}_{fc}^\tau), \tau \in \{a, v\} \quad (2)$$

$$\mathcal{S}_{emb} = \textbf{Embedding}(S, \mathcal{N}_S) \quad (3)$$

$$\mathbf{X}_\tau = \eta \mathcal{S}_{emb} + \mathcal{X}_\tau, \tau \in \{a, v, l\} \quad (4)$$

Method



$$\mathcal{R}_{multi} = \begin{cases} \{(u_i^a, u_i^v), (u_i^v, u_i^a), (u_i^a, u_i^a)\} \\ \{(u_i^v, u_i^l), (u_i^l, u_i^v), (u_i^v, u_i^v)\} \\ \{(u_i^l, u_i^a), (u_i^a, u_i^l), (u_i^l, u_i^l)\} \end{cases} \quad (5)$$

$$\mathcal{R}_{temp} = \begin{cases} \{(u_j \overset{\text{past}}{\rightarrow} u_i)^\tau | i - \mathcal{P} < j < i\} \\ \{(u_i \overset{\text{future}}{\leftarrow} u_j)^\tau | i < j < i + \mathcal{F}\} \end{cases} \quad (6)$$

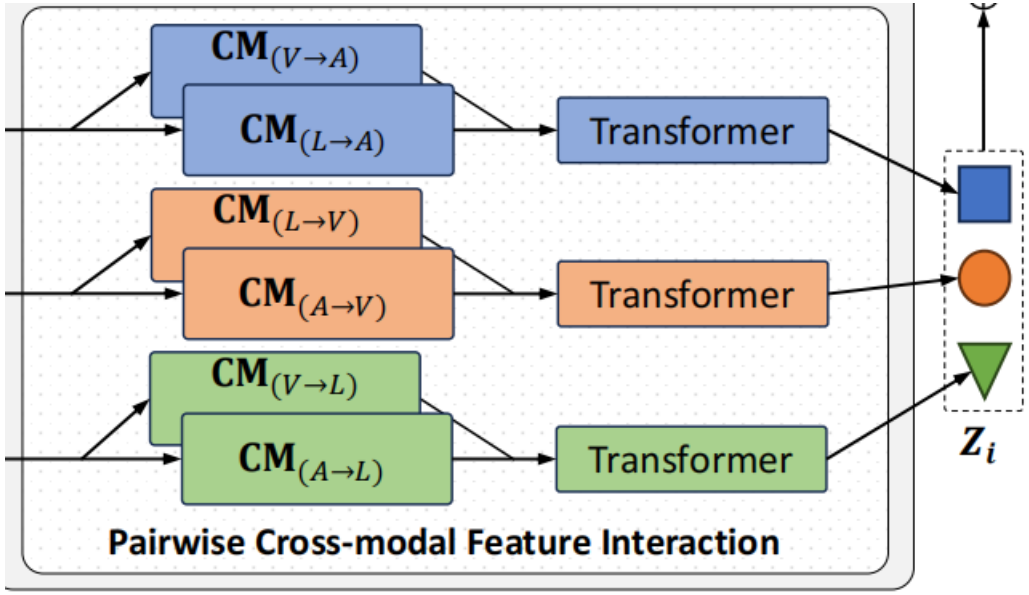
$$g_i^\tau = \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_r(i)} \frac{1}{|\mathcal{N}_r(i)|} \mathbf{W}_r \cdot x_i^\tau + \mathbf{W}_0 \cdot x_i^\tau \quad (7)$$

$$o_i^\tau = ||_{c=1}^{C} [\mathbf{W}_1 g_i^\tau + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j}^\tau \mathbf{W}_2 g_j^\tau] \quad (8)$$

$$\alpha_{i,j}^\tau = softmax(\frac{(\mathbf{W}_3 g_i^\tau)^\top (\mathbf{W}_4 g_i^\tau)}{\sqrt{d}}) \quad (9)$$

$$\mathbf{G}^\tau = \{o_1^\tau, o_2^\tau, \ldots, o_N^\tau\} \quad (10)$$

Method

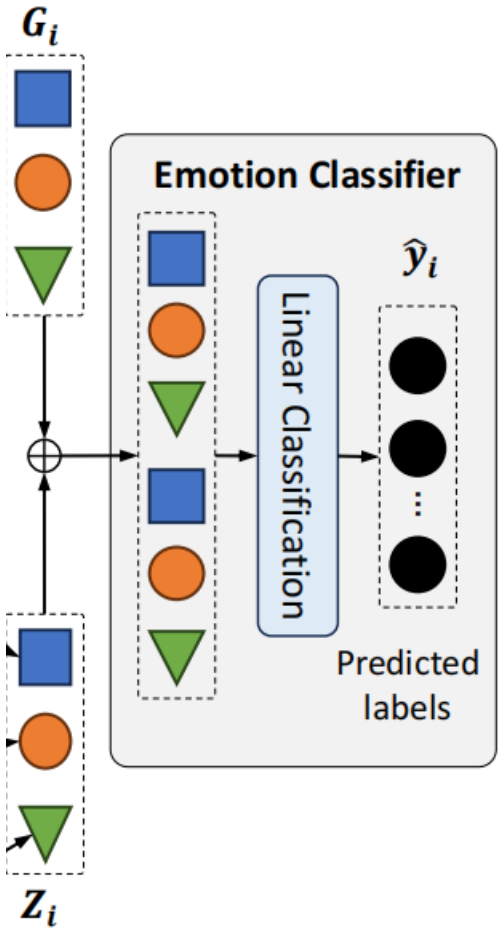


$$\mathbf{CM}^{l\to a} = \sigma\left(\frac{\mathbf{X}^a\mathbf{W}_{Q^a}(\mathbf{W}_{K^l})^\top(\mathbf{X}^l)^\top}{\sqrt{d_k}}\right)\mathbf{X}_l\mathbf{W}_{V^l} \tag{11}$$

$$\begin{aligned}
\mathbf{Z}^{l\to a}_{[0]} &= \mathbf{Z}^{a}_{[0]} \\
\overline{\mathbf{Z}}^{l\to a}_{[i]} &= \mathbf{CM}^{l\to a}_{[i]}(LN(\mathbf{Z}^{l\to a}_{[i-1]}), LN(\mathbf{Z}^{l\to a}_{0})) \\
&\quad + LN(\mathbf{Z}^{l\to a}_{[i-1]}) \\
\mathbf{Z}^{l\to a}_{[i]} &= (LN(\overline{\mathbf{Z}}^{l\to a}_{[i]}))^{FFN} + LN(\overline{\mathbf{Z}}^{l\to a}_{[i]})
\end{aligned} \tag{12}$$

$$\begin{aligned}
LN(\overline{\mathbf{Z}}^{l\to a}_{[i]}))^{FFN} &= \max(0, LN(\overline{\mathbf{Z}}^{l\to a}_{[i]}))\mathbf{\Omega}_1 \\
&\quad + \mathbf{b}_1)\mathbf{\Omega}_2 + \mathbf{b}_2
\end{aligned} \tag{13}$$

Method



$$\mathbf{H} = Fusion([\mathbf{G}, \mathbf{Z}]) \tag{14}$$
$$= [\{o_1^\tau, o_2^\tau, \ldots, o_N^\tau\}, \{\mathbf{Z}_{a\rightleftarrows v}^{[D]}, \mathbf{Z}_{v\rightleftarrows l}^{[D]}, \mathbf{Z}_{l\rightleftarrows a}^{[D]}\}]$$

$$v_i = ReLU(\mathbf{\Phi}_0 h_i + b_0) \tag{15}$$
$$p_i = \text{softmax}(\mathbf{\Phi}_1 v_i + b_1) \tag{16}$$
$$\hat{y}^i = \text{argmax}(p_i) \tag{17}$$

Experiments

Dataset	Dialogues			Utterances		
	train	valid	test	train	valid	test
IEMOCAP (6-way)	108	12	31	5,146	664	1,623
IEMOCAP (4-way)	108	12	31	3,200	400	943
MOSEI	2,249	300	646	16,327	1,871	4,662

Experiments

Methods	IEMOCAP (6-way)							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc. (%)	w-F1 (%)
bc-LSTM (Poria et al., 2017)	32.63	70.34	51.14	63.44	67.91	61.06	59.58	59.10
CMN (Hazarika et al., 2018b)	30.38	62.41	52.39	59.83	60.25	60.69	56.56	56.13
ICON (Hazarika et al., 2018a)	29.91	64.57	57.38	63.04	63.42	60.81	59.09	58.54
DialogueRNN (Majumder et al., 2019)	33.18	78.80	59.21	65.28	71.86	58.91	63.40	62.75
DialogueGCN (Ghosal et al., 2019)	47.10	**80.88**	58.71	66.08	70.97	61.21	65.54	65.04
MMGCN (Wei et al., 2019)	45.45	77.53	61.99	66.70	72.04	64.12	65.56	65.71
DialogueCRN (Hu et al., 2021)	51.59	74.54	62.38	67.25	73.96	59.97	65.31	65.34
COGMEN (Joshi et al., 2022)	55.76	80.17	63.21	61.69	**74.91**	63.90	67.04	67.27
CORECT (Ours)	**59.30**	80.53	**66.94**	**69.59**	72.69	**68.50**	**69.93** (↑ 2.89)	**70.02** (↑ 2.75)

Experiments

Modality Settings	IEMOCAP (4-way)	
	Acc. (%)	w-F1 (%)
bc-LSTM (Poria et al., 2017)	75.20	75.13
CHFusion (Majumder et al., 2018)	76.59	76.80
COGMEN (Joshi et al., 2022)	<u>82.29</u>	<u>82.15</u>
CORECT (Ours)	**84.73** (↑ 2.44)	**84.64** (↑ 2.49)

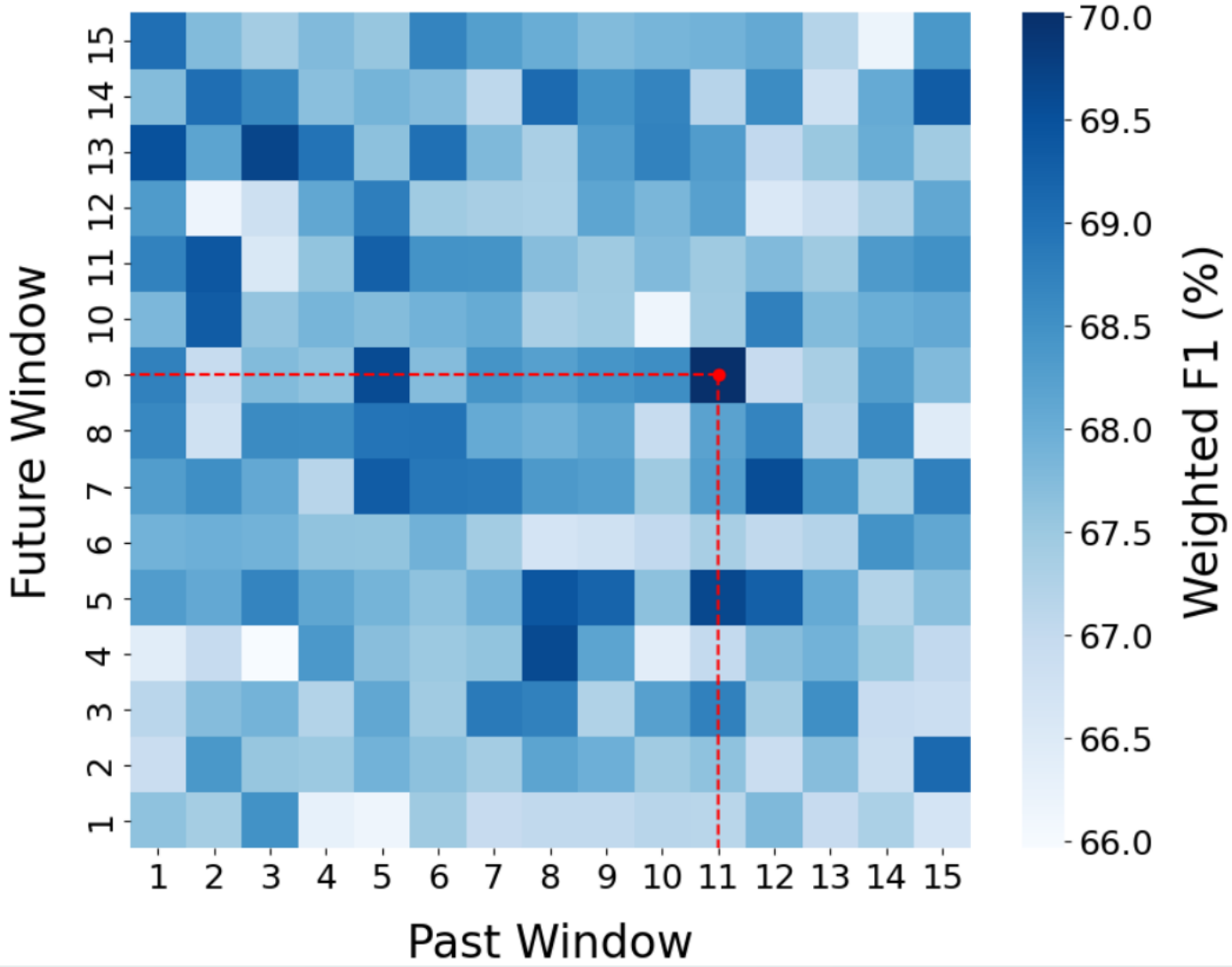
Experiments

Methods	Sentiment Classification Accuracy (%)		Emotion Classification (Binary, 1 vs. all) weighted F1-score (%)					
	2 Class	7 Class	Happiness	Sadness	Angry	Fear	Disgust	Surprise
Multilouge-Net (Shenoy and Sardana, 2020)	82.88	44.83	67.84	65.34	67.03	*87.79*	74.91	*86.05*
TBJE (Delbrouck et al., 2020)	82.40	43.91	65.91	70.78	70.86	*87.79*	82.57	86.04
COGMEN (Joshi et al., 2022)	82.95	45.22	70.88	70.91	74.20	*87.79*	81.83	*86.05*
CORECT (Ours)	**83.66**	**46.31**	**71.35**	**72.86**	**76.77**	**87.90**	**84.26**	**86.48**

Experiments

Modality Settings	IEMOCAP (6-way)		IEMOCAP (4-way)	
	Acc. (%)	w-F1 (%)	Acc. (%)	w-F1 (%)
A	52.31	51.49	67.02	65.48
T	67.22	67.26	82.82	82.65
V	38.63	37.67	49.73	47.97
A+T	68.27	68.36	83.14	83.13
T+V	65.50	65.61	81.76	81.75
V+A	54.16	53.82	69.03	68.21
CORECT (A+T+V)	**69.93**	**70.02**	**84.73**	**84.64**

Experiments



Thanks!