MM-DFN: Multimodal Dynamic Fusion Network for Emtion Recognition in Conversations

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

Dou Hu[1] *Xiaolong Hou*[1] *Lingwei Wei*[2] *Lianxin Jiang*[1] *Yang Mo*[1]

[1] Ping An Life Insurance Company of China, Ltd.
[2] Institute of Information Engineering, Chinese Academy of Sciences

https://github.com/zerohd4869/MM-DFN

ICASSP-2022

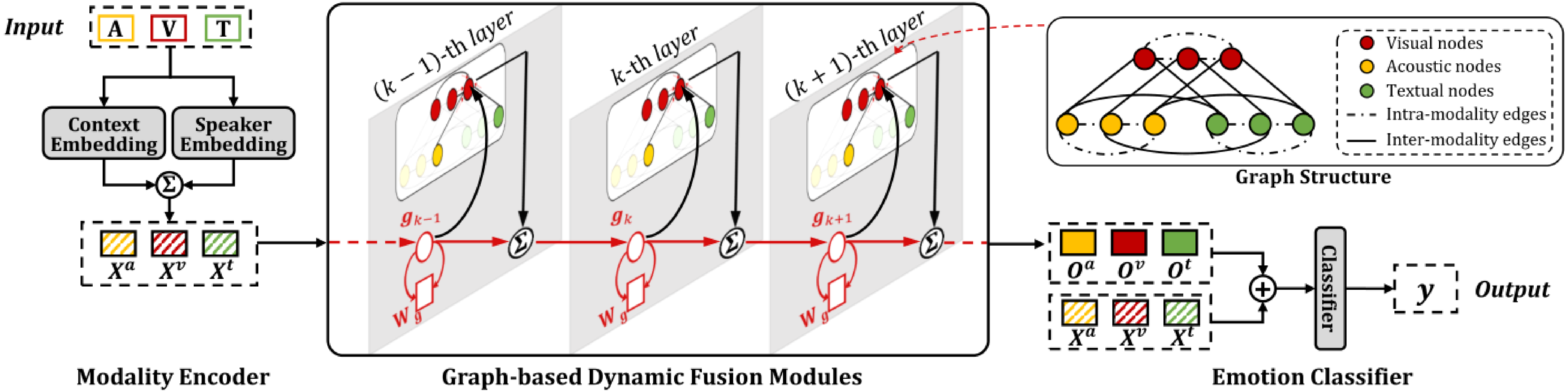
Reported by Jiawei Cheng

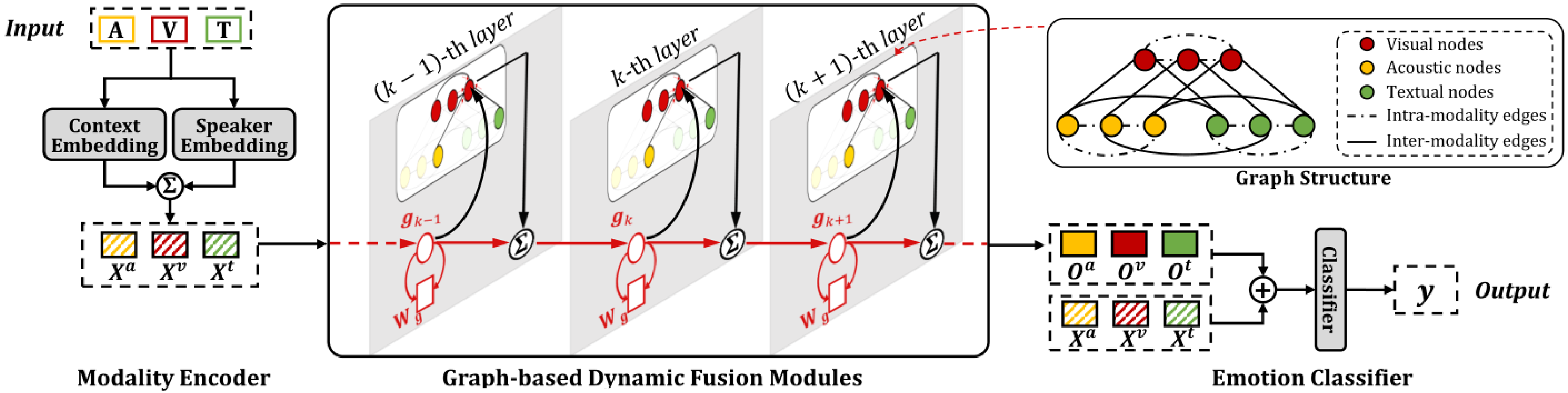
Introduction

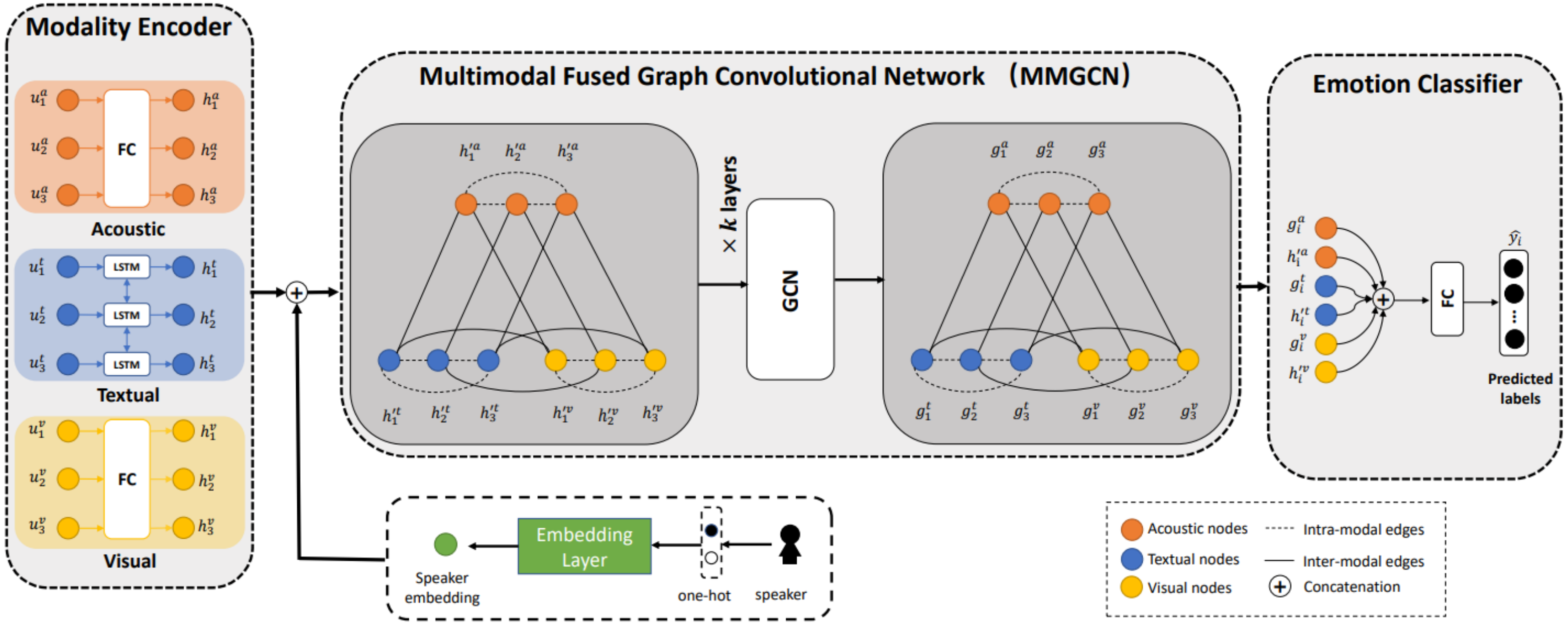
Previous methods ignore complex interactions between utterances, resulting in leveraging context information inconversations insufficiently

However, these graph-based fusion methods aggregate contextual information in a specific semantic space at each layer, gradually accumulating redundant information

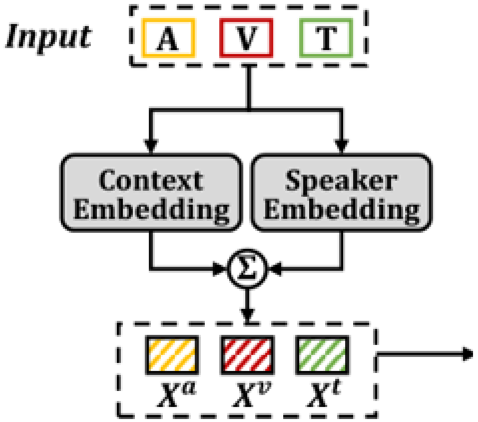
Overview


Input
A
V
T
Context Embedding
Speaker Embedding
Σ
X^a
X^v
X^t
(k − 1)-th layer
k-th layer
(k + 1)-th layer
g_{k-1}
g_k
g_{k+1}
W g
Visual nodes
Acoustic nodes
Textual nodes
Intra-modality edges
Inter-modality edges
Graph Structure
O^a
O^v
O^t
Classifier
y
Output
Modality Encoder
Graph-based Dynamic Fusion Modules
Emotion Classifier


Input
A
V
T
Context Embedding
Speaker Embedding
(k − 1)-th layer
k-th layer
(k + 1)-th layer
Visual nodes
Acoustic nodes
Textual nodes
Intra-modality edges
Inter-modality edges
Graph Structure
Classifier
y
Output
Modality Encoder
Graph-based Dynamic Fusion Modules
Emotion Classifier


Modality Encoder
FC
Acoustic
LSTM
Textual
FC
Visual
Multimodal Fused Graph Convolutional Network （MMGCN）
× k layers
GCN
Emotion Classifier
FC
Predicted labels
Speaker embedding
Embedding Layer
one-hot
speaker
Acoustic nodes
Textual nodes
Visual nodes
Intra-modal edges
Inter-modal edges
Concatenation

Method

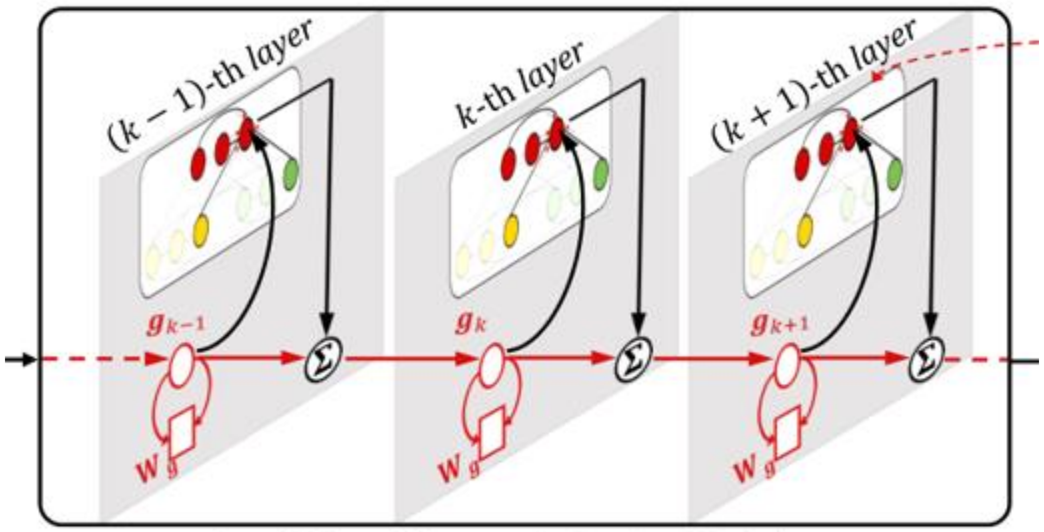


Modality Encoder

$$\mathbf{c}_i^{\varsigma} = \mathbf{W}_c^{\varsigma}\mathbf{u}_i^{\varsigma} + \mathbf{b}_c^{\varsigma}, \varsigma \in \{a, v\},$$
$$\mathbf{c}_i^{t}, \mathbf{h}_i^{c} = \overleftrightarrow{GRU}_c(\mathbf{u}_i^{t}, \mathbf{h}_{i-1}^{c}), \quad (1)$$

$$\mathbf{s}_i^{\delta}, \mathbf{h}_{\lambda,j}^{s} = \overleftrightarrow{GRU}_s(\mathbf{u}_i^{\delta}, \mathbf{h}_{\lambda,j-1}^{s}), j \in [1, |U_\lambda|], \delta \in \{a, v, t\}, \quad (2)$$

Method

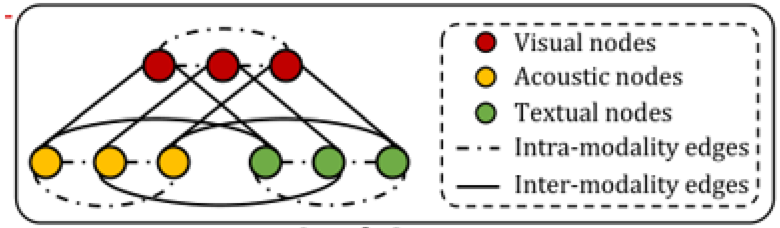


Graph-based Dynamic Fusion Modules

$$\mathbf{x}_i^\delta = \mathbf{c}_i^\delta + \gamma^\delta \mathbf{s}_i^\delta, \delta \in \{a, v, t\}, \tag{3}$$

$$\mathbf{A}_{ij} = 1 - \frac{\arccos(\mathrm{sim}(\mathbf{x}_i, \mathbf{x}_j))}{\pi}$$

$$\begin{aligned}
\mathbf{\Gamma}_\varepsilon^{(k)} &= \sigma(\mathbf{W}_\varepsilon^g \cdot [\mathbf{g}^{(k-1)}, \mathbf{H'}^{(k-1)}] + \mathbf{b}_\varepsilon^g), \varepsilon = \{u, f, o\}, \\
\tilde{\mathbf{C}}^{(k)} &= \tanh(\mathbf{W}_C^g \cdot [\mathbf{g}^{(k-1)}, \mathbf{H'}^{(k-1)}] + \mathbf{b}_C^g), \\
\mathbf{C}^{(k)} &= \mathbf{\Gamma}_f^{(k)} \odot \mathbf{C}^{(k-1)} + \mathbf{\Gamma}_u^{(k)} \odot \tilde{\mathbf{C}}^{(k)}, \ \mathbf{g}^{(k)} = \mathbf{\Gamma}_o^{(k)} \odot \tanh(\mathbf{C}^{(k)}),
\end{aligned} \tag{4}$$



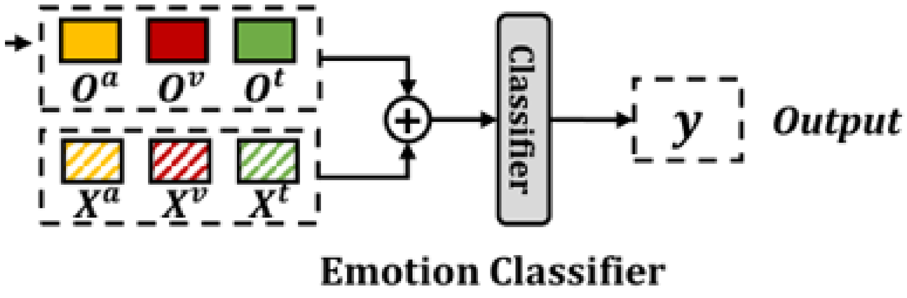
Graph Structure

$$\mathbf{H}^{(k)} = \mathrm{ReLU}\left(((1-\alpha)\tilde{\mathbf{P}}\mathbf{H'}^{(k-1)} + \alpha\mathbf{H}^{(0)})((1-\beta_{k-1})\mathbf{I}_n + \beta_{k-1}\mathbf{W}^{(k-1)})\right), \tag{5}$$

$\beta_k = log(\frac{\rho}{k} + 1)$. $\mathbf{H}^{(0)}$ is initialized with $\mathbf{X}^a, \mathbf{X}^v, \mathbf{X}^t$.

$\mathbf{I}_n$ is an identity mapping matrix $\mathbf{H'}^{(k)} = \mathbf{H}^{(k)} + \mathbf{g}^{(k)}$

Method



$$\hat{\mathbf{y}}_i = \mathrm{Softmax}(\mathbf{W}_z[\mathbf{x}_i^a; \mathbf{x}_i^v; \mathbf{x}_i^t; \mathbf{o}_i^a; \mathbf{o}_i^v; \mathbf{o}_i^t] + \mathbf{b}_z), \quad (6)$$

$$\mathcal{L} = -\frac{1}{\sum_{l=1}^{L} \tau(l)} \sum_{i=1}^{L} \sum_{j=1}^{\tau(i)} \mathbf{y}_{i,j}^l log(\hat{\mathbf{y}}_{i,j}^l) + \eta \|\Theta\|_2, \quad (7)$$

Experiments

Methods	IEMOCAP								MELD						
	Happy	*Sad*	*Neutral*	*Angry*	*Excited*	*Frustrated*	Acc	w-F1	*Neutral*	*Surprise*	*Sadness*	*Happy*	*Anger*	Acc	w-F1
TFN [9]	37.26	65.21	51.03	54.64	58.75	56.98	55.02	55.13	77.43	47.89	18.06	51.28	44.15	60.77	57.74
LMF [10]	37.76	66.53	52.39	57.53	58.41	59.27	56.50	56.49	76.97	47.06	21.15	54.20	46.64	61.15	58.30
MFN [11]	48.19	73.41	56.28	63.04	64.11	61.82	61.24	61.60	77.27	48.29	23.24	52.63	41.32	60.80	57.80
bc-LSTM [6]	33.82	78.76	56.75	64.35	60.25	60.75	60.51	60.42	75.66	48.57	22.06	52.10	44.39	59.62	57.29
ICON [7]	32.80	74.40	60.60	68.20	68.40	66.20	64.00	63.50	-	-	-	-	-	-	-
DialogueRNN [4]	32.20	80.26	57.89	62.82	73.87	59.76	63.52	62.89	76.97	47.69	20.41	50.92	45.52	60.31	57.66
DialogueCRN [3]	**53.23**	**83.37**	62.96	66.09	75.40	66.07	67.16	67.21	77.01	50.10	26.63	52.77	45.15	61.11	58.67
DialogueGCN [5]	51.57	80.48	57.69	53.95	72.81	57.33	63.22	62.89	75.97	46.05	19.60	51.20	40.83	58.62	56.36
MMGCN [13]	45.14	77.16	64.36	68.82	74.71	61.40	66.36	66.26	76.33	48.15	**26.74**	53.02	46.09	60.42	58.31
MM-DFN	42.22	78.98	**66.42***	**69.77***	**75.56***	**66.33***	**68.21***	**68.18***	**77.76***	**50.69***	22.93	**54.78***	**47.82***	**62.49***	**59.46***

Table 1. Results under the multimodal setting (A+V+T). We present the overall performance of Acc and w-F1, which mean the overall accuracy score and weighted-average F1 score, respectively. We also report F1 score per class, except two classes (i.e. *Fear* and *Disgust*) on MELD, whose results are not statistically significant due to the smaller number of training samples. Best results are highlighted in bold. * represents statistical significance over state-of-the-art scores under the paired-t test ($p < 0.05$).

Experiments

Methods	IEMOCAP	MELD
MM-DFN	68.18	59.46
- w/o GDF - w Speaker - w Context	63.80	58.50
- w GDF - w/o Speaker - w Context	66.89	58.45
- w/o GDF - w/o Speaker - w Context	62.90	58.50
- w/o GDF - w/o Speaker - w/o Context	54.81	58.08

Table 2. Ablation results of MM-DFN. We report w-F1 score for both datasets.

Experiments

Fusion Modules	IEMOCAP	MELD
Concat / Gate Fusion	63.80 / 64.30	58.50 / 57.87
Tensor / Memory Fusion	61.05 / 65.51	58.54 / 58.48
Early / Late Fusion + GCN	64.19 / 65.34	58.69 / 58.43
Graph-based Fusion (GF)	67.02	58.54
- w/o Inter-Modal - w Intra-Modal	66.91	58.53
- w Inter-Modal - w/o Intra-Modal	66.11	58.29
Graph-based Dynamic Fusion (GDF)	68.18	59.46
- w/o Inter-Modal - w Intra-Modal	67.82	59.15
- w Inter-Modal - w/o Intra-Modal	66.22	58.31

Table 3. Results against different fusion modules. We report w-F1 score for both datasets.

Experiments

Modality	IEMOCAP		MELD	
	GF	GDF	GF	GDF
A / V / T	-	47.79 / 27.46 / 61.07	-	42.72 / 32.34 / 56.95
A + V	54.73	56.35	42.74	44.67
A + T	65.03	65.41	57.85	58.34
V + T	62.07	62.63	57.78	58.49
A + V + T	67.02	68.18	58.54	59.46

Table 4. Results of graph-based fusion methods under different modality settings. Fusion modules are not used under unimodal types. We report w-F1 score for both datasets.

Thanks