Multivariate, Multi-Frequency and Multimodal: Rethinking Graph Neural Networks for Emotion Recognition in Conversation (Appendix)

Feiyu Chen[†‡] Jie Shao[†‡] Shuyuan Zhu[†] Heng Tao Shen[†‡]
[†]University of Electronic Science and Technology of China, Chengdu, China
[‡]Sichuan Artificial Intelligence Research Institute, Yibin, China
{chenfeiyu,shaojie,eezsy,shenhengtao}@uestc.edu.cn

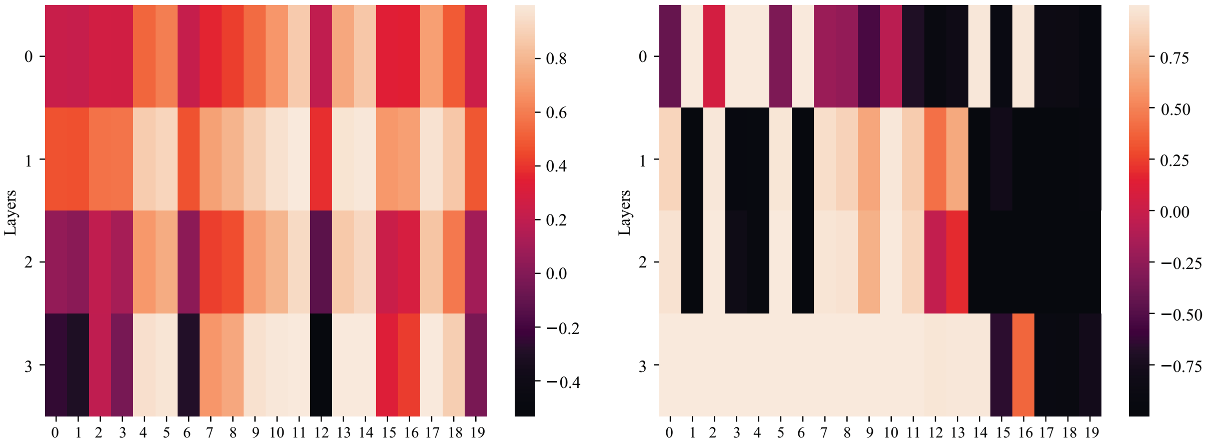

Figure 5. Visualization examples of the coefficient $r^l_{ij} - r^h_{ij}$ at different layers k on IEMOCAP.

A. Unimodal feature extraction

The textual features are extracted using the RoBERTa Large model [4], which is firstly fine-tuned for emotion prediction from the transcript of conversations. After the fine-tuning process, the utterances are fed to the model and the activations from the final four layers are extracted as four textual vectors, which are then normalized and averaged for the final textual representation. The dimension of textual features in our paper is 1024.

The acoustic features are obtained by the openSMILE toolkit [2]. $IS13$ configuration file is used for the IEMOCAP dataset and $IS10$ configuration file is used for the MELD dataset.

The visual features are extracted with a DenseNet [3] pre-trained on the Facial Expression Recognition Plus (FER+) corpus [1] for the MELD dataset. For the IEMOCAP dataset, 3D-CNN is used with feature maps being 128 for 3D filters of size 5, followed by a max-pooling operation and an activation.

B. Visualization

In Section 3.3.3, we introduce a coefficient $r^l_{ij} - r^h_{ij}$ to model the varying importance of different frequency constituents. If $r^l_{ij} - r^h_{ij} < 0$, the high-frequency messages dominate, and vice versa. Figure 5 shows two visualization examples of the coefficients of two different nodes at different layers on IEMOCAP. It can be observed that, the coefficient allows each central node to adaptively receive messages in different frequency from different neighbours. The patterns also vary across layers, leading to better flexibility for capturing emotion discrepancy and commonality.

C. Additional experimental results

In this section, we provide additional experimental results, including the results of each individual label and the detailed complexity analysis. Table 4 shows the results of each individual label and M^{3}Net outperforms prior works in the majority of classes. Table 5 presents the comparison of computational complexity and shows that M^{3}Net can boost performance without extra burden.

Methods	IEMOCAP								MELD								
	Happy	*Sad*	*Neutral*	*Angry*	*Excited*	*Frustrated*	Acc.	F1	*Neutral*	*Surprise*	*Fear*	*Sad*	*Joy*	*Disgust*	*Anger*	Acc.	F1
DialogueGCN	57.89	77.89	65.51	64.72	68.34	57.42	63.96	64.44	77.79	57.29	7.69	36.66	62.77	16.07	45.49	63.49	62.78
MMGCN	52.70	81.45	62.66	67.79	73.28	61.83	66.79	66.99	78.98	59.47	13.33	40.00	63.04	15.19	54.29	66.63	65.13
DialogueRNN	52.44	80.87	70.53	67.00	69.20	62.76	68.64	68.72	78.89	58.78	**28.30**	40.35	63.50	26.28	53.28	65.94	65.31
MetaDrop	**63.98**	81.67	69.80	62.97	72.66	64.28	69.38	69.59	79.17	59.52	24.78	40.88	64.48	28.57	54.60	66.63	66.30
MM-DFN	46.22	83.37	**71.15**	68.21	75.09	63.82	69.87	69.48	79.70	58.78	20.22	**41.98**	62.90	29.13	53.99	67.01	66.17
M^{3}Net (ours)	62.05	**83.67**	70.80	**68.67**	**79.66**	**64.43**	**72.46**	**72.49**	**80.13**	**60.25**	26.32	39.64	**64.61**	**29.80**	**56.54**	**68.28**	**67.05**

Table 4. Results of each considered primary emotion classification task (F1 score per category).

Methods	DialogueGCN	MMGCN	DialogueRNN	MetaDrop	MM-DFN	M^{3}Net (ours)
FLOPs (I)	10.86G	2.29G	19.67G	26.90G	13.66G	5.18G
Params (I)	3.92M	2.64M	4.21M	9.41M	2.27M	3.50M
FLOPs (M)	1.81G	1.09G	10.08G	15.93G	10.40G	1.24G
Params (M)	2.51M	1.50M	3.56M	6.17M	1.17M	3.52M

Table 5. Comparison of computational complexity. (I) refers to results on IEMOCAP, and (M) refers to results on MELD.

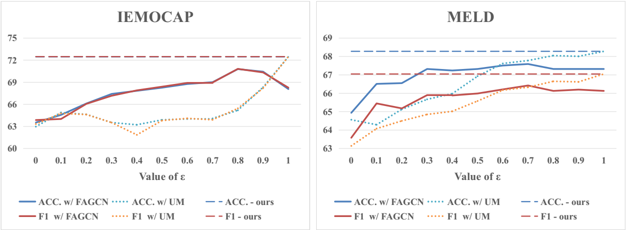

Figure 6. Comparison with FAGCN (solid lines) and replacing the updating mode only (dotted lines). UM = updating mode.

Moreover, as we discussed in Section 3.3.4, in our multi-frequency module, in addition to the hyper-parameter, the node embedding updating mode is another important distinction from FAGCN. Hence, we further conduct experiments to verify whether using this embedding updating mode only can produce good results. We replace the updating mode in FAGCN with the one in Eq. (10) and conduct experiments under different values of ϵ. The results are shown in Figure 6. It can be seen that the results are inferior to ours, and to FAGCN in most cases, which further suggests the necessity and superiority of our design.

References

[1] Emad Barsoum, Cha Zhang, Cristian Canton-Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016*, pages 279–283, 2016. 1

[2] Florian Eyben, Martin Wöllmer, and Björn W. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th International Conference on Multimedia 2010*, pages 1459–1462, 2010. 1

[3] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 2261–2269, 2017. 1

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 1