Multimodal Emotion Recognition Based on Deep Temporal Features Using Cross-modal Transformer and Self-Attention

Paper ID- 4629

SiliconTech

Bubai Maji[1], Monorama Swain[2], Rajlakshmi Guha[1], and Aurobinda Routray[1]

[1]Indian Institute of Technology Kharagpur, [2]Silicon Institute of Technology Bhubaneswar, India

IIT Kharagpur

All the implementation codes are available at: https://github.com/bubaimaji/cmt-mser

4-10th June,
IEEE ICASSP 2023, Greece

I. INTRODUCTION

❑ **Nowadays, multimodal speech emotion recognition (SER) has received more attention due to fusing multimodal information such as audio, text and visual**

- Recent SER studies achieved high accuracy; however, the speakers emotional state is not fully understood
- Selecting large number hand-crafted features are required for better performance
- In this work, a deep learning-based multimodal SER has been proposed

II. MOTIVATION

- The interactive relations between different modalities of speech representations for emotion recognition have not yet been well investigated
- Streaming end-to-end ASER are still lacking success due to low efficacy
- Fusion of high-level features from different modalities becomes a major issue in multimodal emotion recognition tasks

III. CONTRIBUTIONS

❑ **We present a cross-modal Transformer (CMT) and self-attention (SA) based framework for multimodal SER task**

- We used large set (125-dimensions) of hand-crafted features
- A CMT block is designed to capture better inter- and intra-interactions and temporal information between the audio and textual features
- Then the SA network is employed to utilize weighted emotional information from the fused multimodal features to improve the performance

IV. THE PROPOSED METHOD

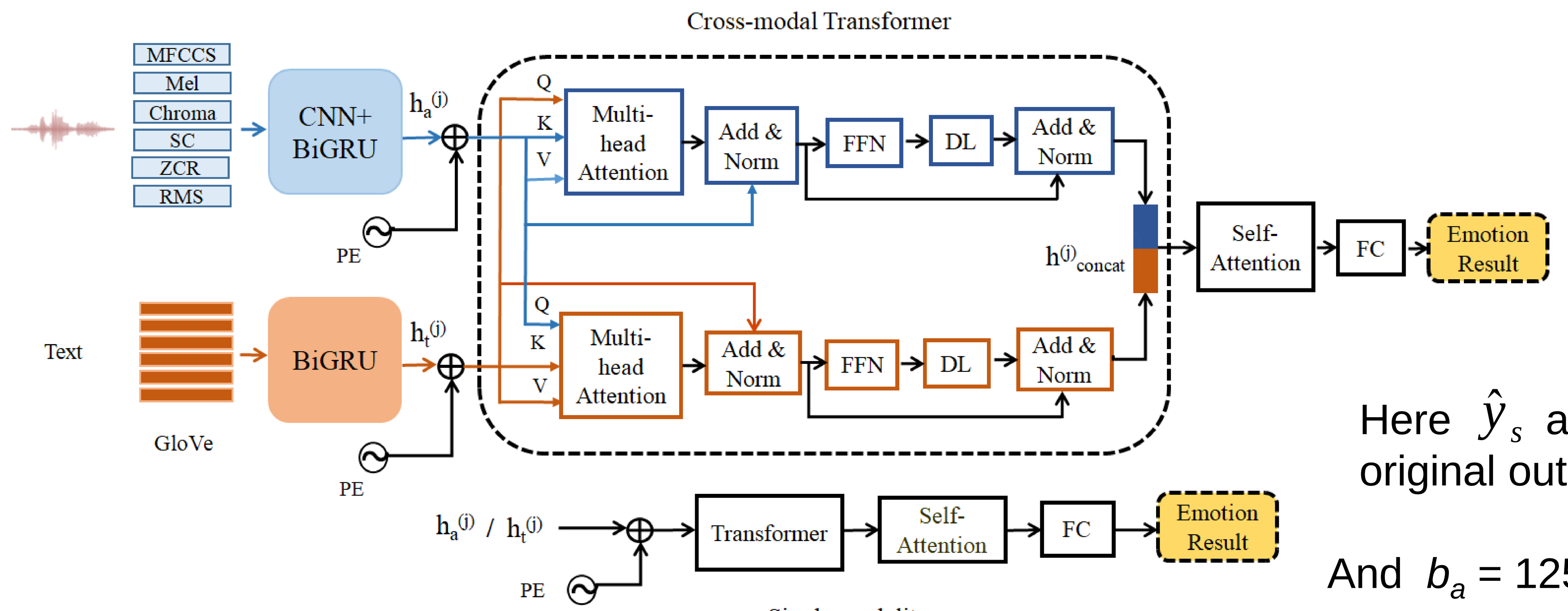

Figure 1: Architecture of the proposed method

Step 1: Audio features are learned by CNN+BiGRU $h_a^{(j)} \in R^{b_a}$

Step 2: Text are represent by Glove vector and by Bi-GRU $h_t^{(j)} \in R^{b_t}$

Step 3: $h_a^{(j)}$ and $h_t^{(j)}$ learned by CMT, represent as $h_{concat}^{(j)} \in R^{2b}$

Model Training

$$f = Relu\,(PW_p + b_p) \quad (1)$$

$$\hat{y} = Softmax\,(W_f + b_f) \quad (2)$$

$$L = -\sum_{s=1}^{S} y_s \log(\hat{y}_s) \quad (3)$$

Here $\hat{y}_s$ and y_s are the predict and original output of the class

And b_a = 125 dimensions feature vector

Step 4: Output of the CMT is pass through SA and represent as $P_{att}^{(j)} \in R^{b2}$

Step 5: Then we use a FC and predict emotion using Softmax function

V. EXPERIMENTS

V-I. Experimental Setup

Emotion	Angry	Happy	Neutral	Sad	Total
Number	1103	1636	1708	1084	5531

Table 1: Sample distribution on IEMOCAP

- To compare with previous works [1, 2, 3], we used four emotion classes

❑ **Evaluation**

We adopt 5-fold, 10-fold cross-validation and Session 5 as test techniques

- **Effects of Bi-GRU and Transformer layers on the model**

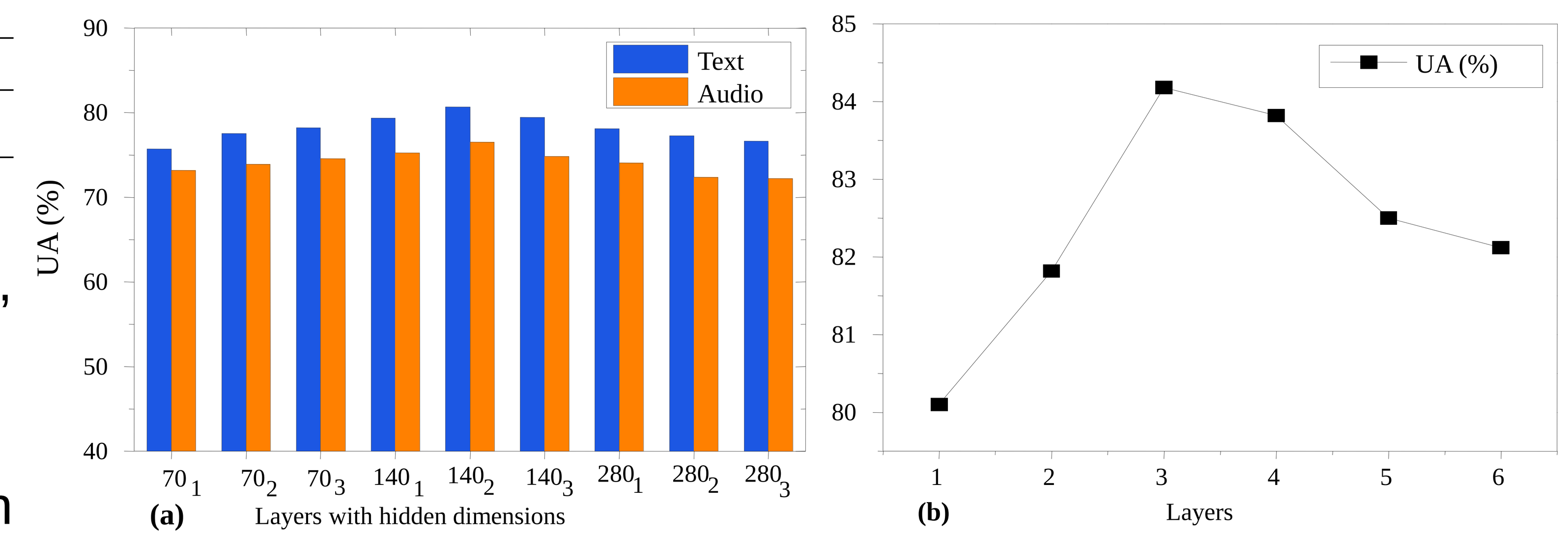

Figure 2: Performances for different (a) hidden dimension with different layers in Bi-GRU (b) number of TLs in CMT

V-II. Evaluation Results

- **Does different cross-validation (CV) cause enhanced model performance?**

# of Fold	Modality	WA (%)	UA (%)
CV-5	A	71.09±0.42	71.84±0.38
CV-5	T	75.18±0.36	76.51±0.55
CV-5	A+T	**78.82±0.50**	**79.95±0.66**
Session 5	A	75.68±0.54	76.85±0.48
Session 5	T	80.13±1.08	80.66±0.73
Session 5	A+T	**83.57±0.71**	**84.43±0.80**
CV-10	A	74.31±0.85	75.69±0.78
CV-10	T	79.81±0.77	80.24±1.21
CV-10	A+T	**80.63±0.90**	**81.49±1.14**

Table 2: The results of the proposed model

Methods	Modality	WA (%)	UA (%)
CV-5			
Liu et al. [5]	A+T	72.39	70.08
Santoso et al. [6]	A+T	76.10	75.90
Makiuchi et al.[3]	A+T	73.50	73
Chen et al. [1]	A+T	74.30	75.30
Wu et al. [2]	A+T	77.57	78.41
Proposed	A+T	**78.82**	**79.95**
CV-10			
Li et al. [7]	A+T	–	79.20
Yoon et al. [4]	A+T	76.50	77.60
Wu et al. [2]	A+T	77.76	78.30
Proposed	A+T	**80.63**	**81.49**
Session 5			
Hu et al. [8]	A+T+V	70.66	70.56
Wu et al. [2]	A+T	83.08	83.22
Proposed	**A+T**	**83.57**	**84.43**

Table 3: Comparison with state-of-the-art methods

V-III. Evaluation Results

❑ **Ablation study**

- Table 4 show the impact of each module in our system
- There is a significant performance reduction when using only a unimodal
- CMT with Bi-GRU and SA perform best among all methods

WA	UA	Mod	Bi-GRU	CMT	SA
69.18	70.20	A		✓	
70.16	70.91	A	✓		✓
72.37	73.05	A		✓	✓
75.68	**76.85**	A	✓	✓	✓
72.34	73.51	T		✓	
71.06	72.45	T	✓		✓
75.26	76.17	T		✓	✓
80.13	**80.66**	T	✓	✓	✓
75.05	76.76	A+T		✓	
77.39	78.21	A+T	✓		✓
80.26	81.64	A+T		✓	✓
83.57	**84.43**	A+T	✓	✓	✓

Table 4: Ablation study of the proposed model

VI. CONCLUSION

- We demonstrate that the transformer alignment network can lead to deeper interaction between different modalities to enhance performance
- The proposed method performs significantly better than the most recent state-of-the-art MSER methods
- Future work: We plan build a real-time application which allows to detect their emotional states automatically

REFERENCES

1 W. Chen, X. Xing, X. Xu, J. Yang, and J. Pang, "Key-Sparse Transformer for Multimodal Speech Emotion Recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6897–6901.

2 W. Wu, C. Zhang, and P. C. Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6269–6273.

3 M. R. Makiuchi, K. Uto, and K. Shinoda, "Multimodal Emotion Recognition with High-Level Speech and Text Features," 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2021, pp. 350–357.