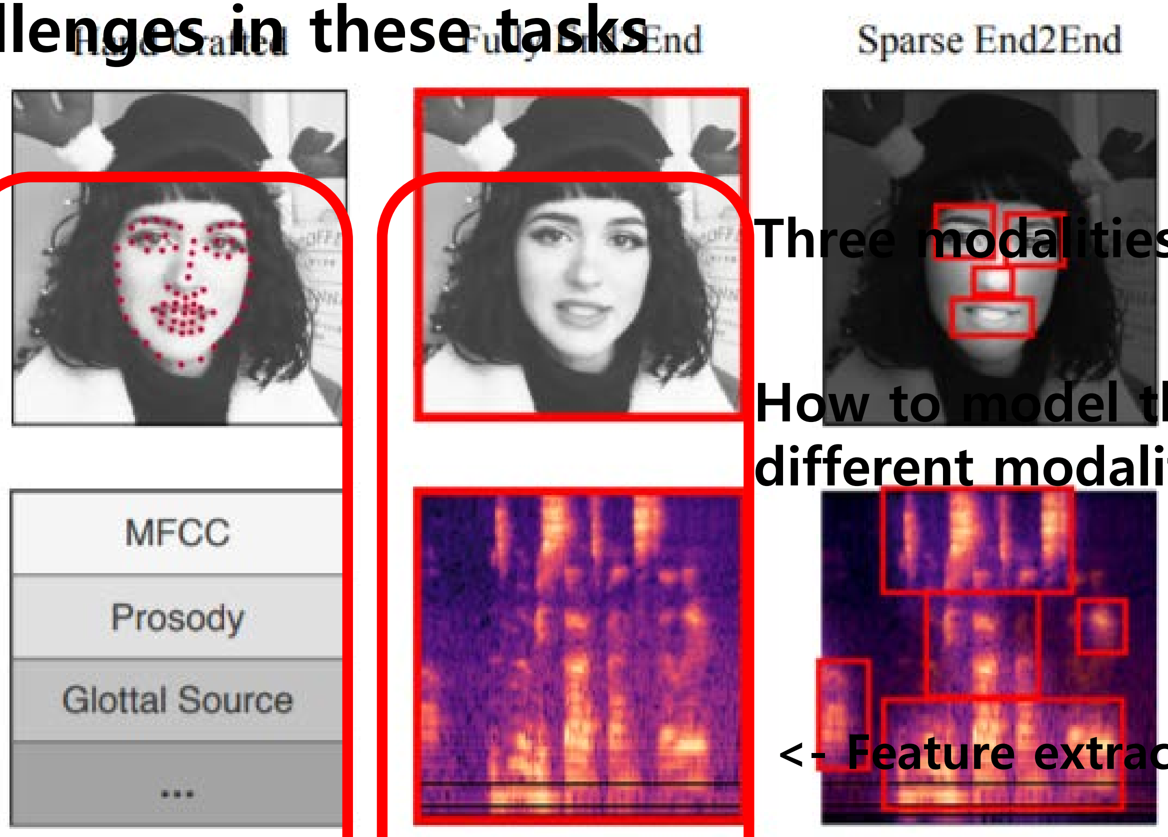
[NAACL 2021] Multimodal End-to-End Sparse Model for Emotion Recognition

Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, Pascale Fung

2021. 11. 22 Eung yeop Kim

1. Abstract & Introduction

The main challenges in these tasks

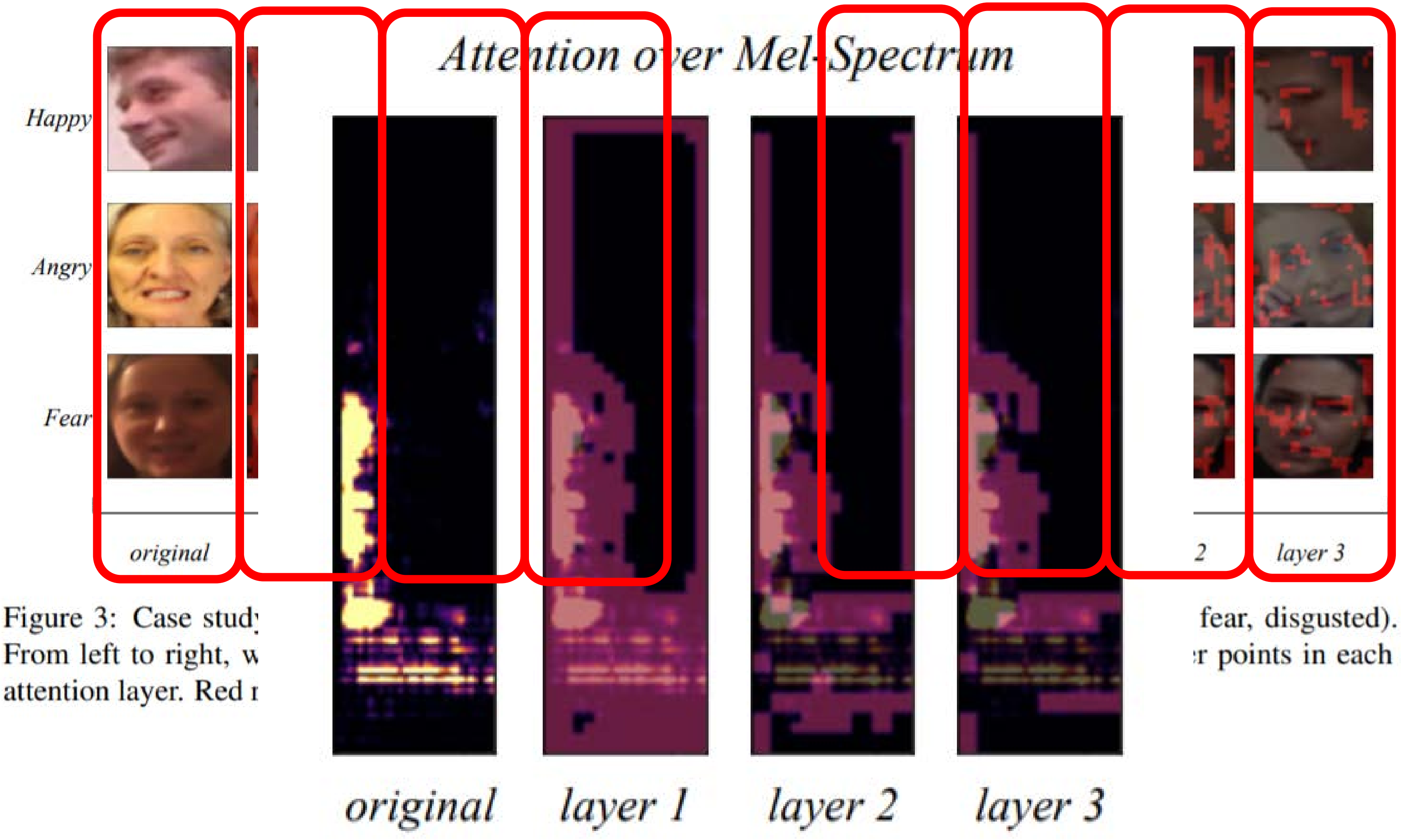


Three modalities : Textual Acoustic Visual

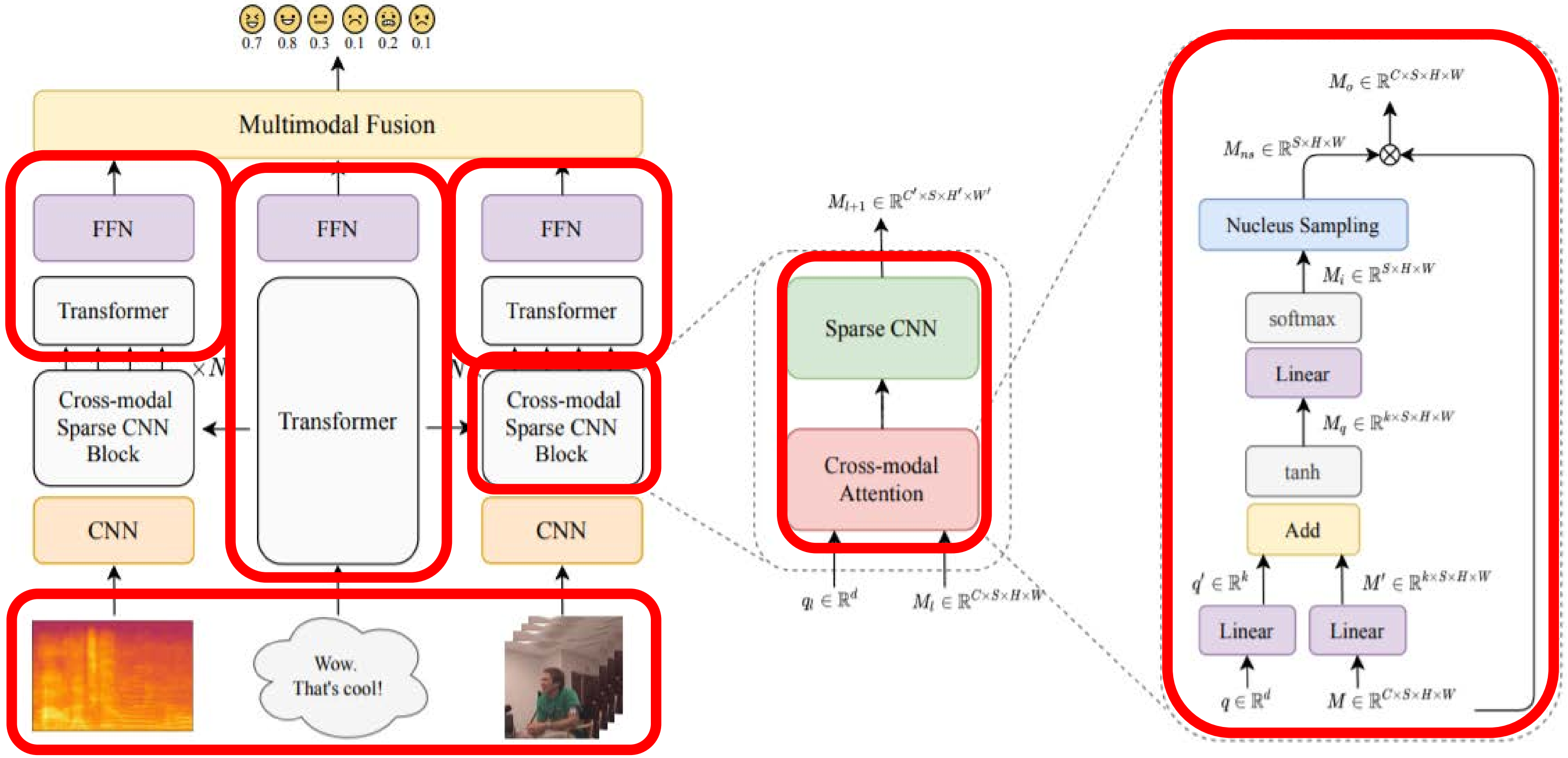
How to model the interactions between different modalities?

<- Feature extractions resulted from each models

1. Abstract & Introduction



2. Methodology



$$M_q = \tanh\left((W_m M + b_m) \oplus W_q q\right) \quad (1)$$
$$M_i = \text{softmax}\left(W_i M_q + b_i\right) \quad (2)$$
$$M_{ns} = \text{Nucleus Sampling}\left(M_i\right) \quad (3)$$
$$M_o = M_{ns} \otimes M, \quad (4)$$

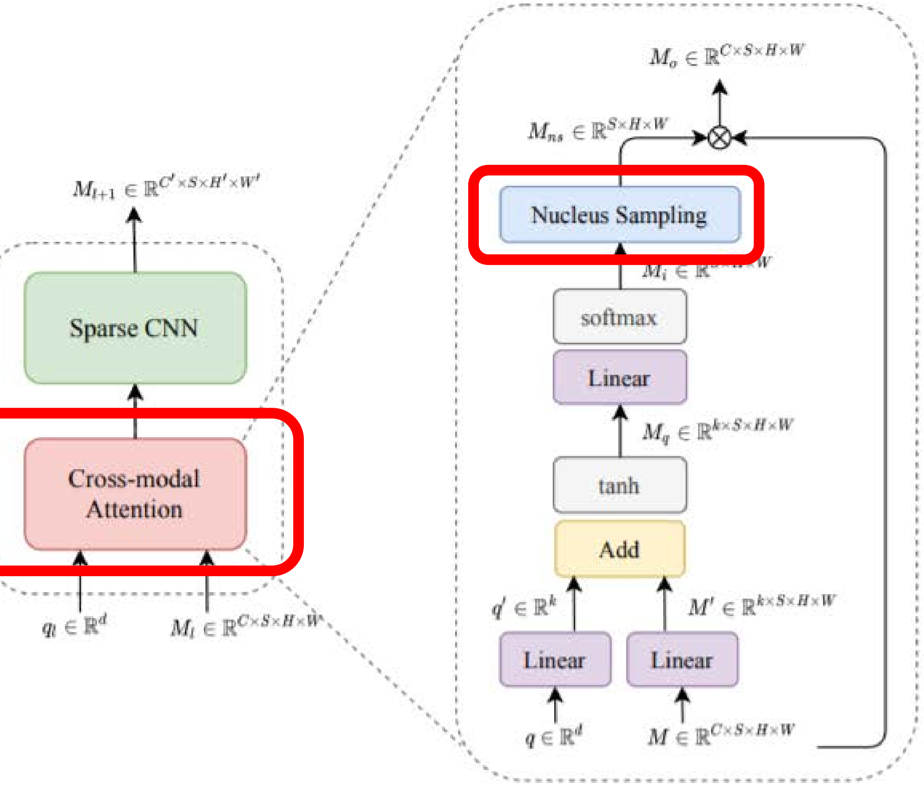
I multimodal data samples : $X = \{(t_i, a_i, v_i)\}_{i=1}^{I}$

t_i is a sequence of words

a_i is a sequence of spectrogram chunks from the audio

v_i is a sequence of RGB image frames

2. Methodology



$$M_q = \tanh\left((W_m M + b_m) \oplus W_q q\right) \quad (1)$$
$$M_i = \text{softmax}\left(W_i M_q + b_i\right) \quad (2)$$
$$M_{ns} = \text{Nucleus Sampling}\left(M_i\right) \quad (3)$$
$$M_o = M_{ns} \otimes M, \quad (4)$$

a query vector $q \in \mathbb{R}^d$

a stack of feature maps $M \in \mathbb{R}^{C\times S\times H\times W}$, where C, S, H, and W

$W_m \in \mathbb{R}^{k\times C}$, $W_q \in \mathbb{R}^{k\times d}$, and $W_i \in \mathbb{R}^k$ are linear transformation weights

the softmax function is applied to the $(H \times W)$ di-
mensions, and $M_i \in \mathbb{R}^{S\times H\times W}$ is the tensor of the
spatial attention scores corresponding to each fea-ture map.

we perform Nucleus Sampling

Therefore, M_o is a sparse tensor with some positions being zero, and the degree of sparsity is controlled by p.

0. Evaluation Metrics

1) Evaluation Metrics

- IEMOCAP dataset : Accuracy, F1-score
- CMU-MOSEI dataset : Weighted Accuracy, F1-score

2) Weighted Accuracy (WAcc) for evaluating the CMU-MOSEI

- It contains many more negative samples than positive ones on each emotion category.
- If normal accuracy is used, a model will still get a fine score when predicting all samples to be neg-.

$$\text{WAcc.} = \frac{TP \times N/P + TN}{2N},$$

in which P means total positive, TP true positive,
N total negative, and TN true negative.

3. Experiments

1. Feature extraction step

Label	Avg. word length	Avg. clip duration (s)	Train size	Valid size	Test size
Anger	15.96	4.51	757	112	234
Excited	16.79	4.78	736	92	213
Frustrated	17.14	4.71	1298	180	371
Happiness	13.58	4.34	398	62	135
Neutral	13.08	3.90	1214	173	321
Sadness	14.82	5.50	759	118	207

Table 1: Statistics of our IEMOCAP dataset split.

Label	Avg. word length	Avg. clip duration (s)	Train size	Valid size	Test size
Anger	7.75	23.24	3267	318	1015
Disgust	7.57	23.54	2738	273	744
Fear	10.04	28.82	1263	169	371
Happiness	8.14	24.12	7587	945	2220
Sadness	8.12	24.07	4026	509	1066
Surprise	8.40	25.95	1465	197	393

Table 2: Statistics of our CMU-MOSEI dataset split.

Visual data : **35 facial action units** using the Open Face library for the image frames in the video, which **capture the movement of facial muscles.**

Acoustic data : a total of 142 dimension features consisting of 12 dimension bark band energy(**BBE**) features, 22 dimension mel-frequency cepstral coefficient(**MFCC**) features, and 108 statistical features from 18 phonological classes. **We extract the features per 400 ms time frame using the DisVoice library.**

Textual data : the pre-trained GloVe word embeddings.(glove.840B.300d)

3. Experiments

Model	#FLOPs ($\times 10^9$)	Angry Acc.	Angry F1	Excited Acc.	Excited F1	Frustrated Acc.	Frustrated F1	Happy Acc.	Happy F1	Neutral Acc.	Neutral F1	Sad Acc.	Sad F1	Average Acc.	Average F1
LF-LSTM	-	71.2	49.4	79.3	57.2	68.2	51.5	67.2	37.6	66.5	47.0	78.2	54.0	71.8	49.5
LF-TRANS	-	81.9	50.7	85.3	57.3	60.5	49.3	85.2	37.6	72.4	49.7	87.4	57.4	78.8	50.3
EmoEmbs†	-	65.9	48.9	73.5	58.3	68.5	52.0	69.6	38.3	73.6	48.7	80.8	53.0	72.0	49.8
MulT†	-	77.9	60.7	76.9	58.0	72.4	57.0	80.0	46.8	74.9	53.7	83.5	65.4	77.6	56.9
FE2E	8.65	**88.7**	**63.9**	**89.1**	**61.9**	71.2	57.8	**90.0**	44.8	**79.1**	**58.4**	**89.1**	**65.7**	**84.5**	**58.8**
MESM ($p = 0.7$)	**5.18**	88.2	62.8	88.3	61.2	**74.9**	**58.4**	89.5	**47.3**	77.0	52.0	88.6	62.2	**84.4**	57.4

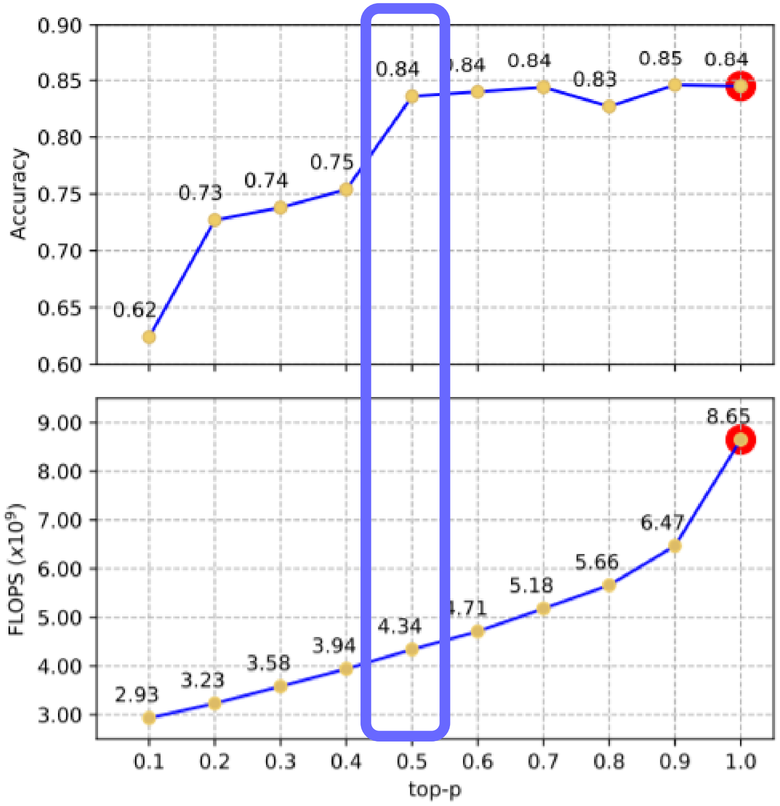
Table 3: The results on the IEMOCAP dataset. #FLOPs is the number of floating point operations per second. We report the accuracy (Acc.) and the F1-score on six emotion categories: *angry*, *excited*, *frustrated*, *happy*, *neutral* and *sad*. We re-run the models marked by †, as we use two more categories and the split is different.

Model	#FLOPs ($\times 10^9$)	Angry WAcc.	Angry F1	Disgusted WAcc.	Disgusted F1	Fear WAcc.	Fear F1	Happy WAcc.	Happy F1	Sad WAcc.	Sad F1	Surprised WAcc.	Surprised F1	Average WAcc.	Average F1
LF-LSTM	-	64.5	47.1	70.5	49.8	61.7	22.2	61.3	73.2	63.4	47.2	57.1	20.6	63.1	43.3
LF-TRANS	-	65.3	47.7	74.4	51.9	62.1	24.0	60.6	72.9	60.1	45.5	62.1	24.2	64.1	44.4
EmoEmbs†	-	66.8	49.4	69.6	48.7	63.8	23.4	61.2	71.9	60.5	47.5	63.3	24.0	64.2	44.2
MulT†	-	64.9	47.5	71.6	49.3	62.9	25.3	**67.2**	**75.4**	64.0	48.3	61.4	25.6	65.4	45.2
FE2E	8.65	**67.0**	**49.6**	**77.7**	**57.1**	63.8	26.8	65.4	72.6	**65.2**	**49.0**	**66.7**	**29.1**	**67.6**	**47.4**
MESM (0.5)	**4.34**	66.8	49.3	75.6	56.4	**65.8**	**28.9**	64.1	72.3	63.0	46.6	65.7	27.2	66.8	46.8

Table 4: The results on the CMU-MOSEI dataset. WAcc stands for weighted accuracy. We report the accuracy and the F1-score on six emotion categories: *angry*, *disgusted*, *fear*, *happy*, *sad* and *surprised*. We re-run the models marked by †, as the data we use is unaligned along the sequence length dimension and the split is different.

3. Experiments & Ablation Study

Effects of Nucleus Sampling



Model	Mods.	Avg. Acc	Avg. F1
	TAV	84.5	58.5
	TA	83.7	54.0
	TV	82.8	55.7
FE2E	VA	81.2	54.4
	T	80.8	50.0
	A	73.3	44.9
	V	78.2	49.8
	TAV	84.4	57.3
MESM	TA	83.6	56.7
	TV	82.1	56.0

T : Textual
A : Acoustic
V : Visual

Specifically, with a top-p of 0.5, the MESM can achieve comparable performance to the FE2E model with around half of the FLOPs in the feature extraction.

3. Experiments

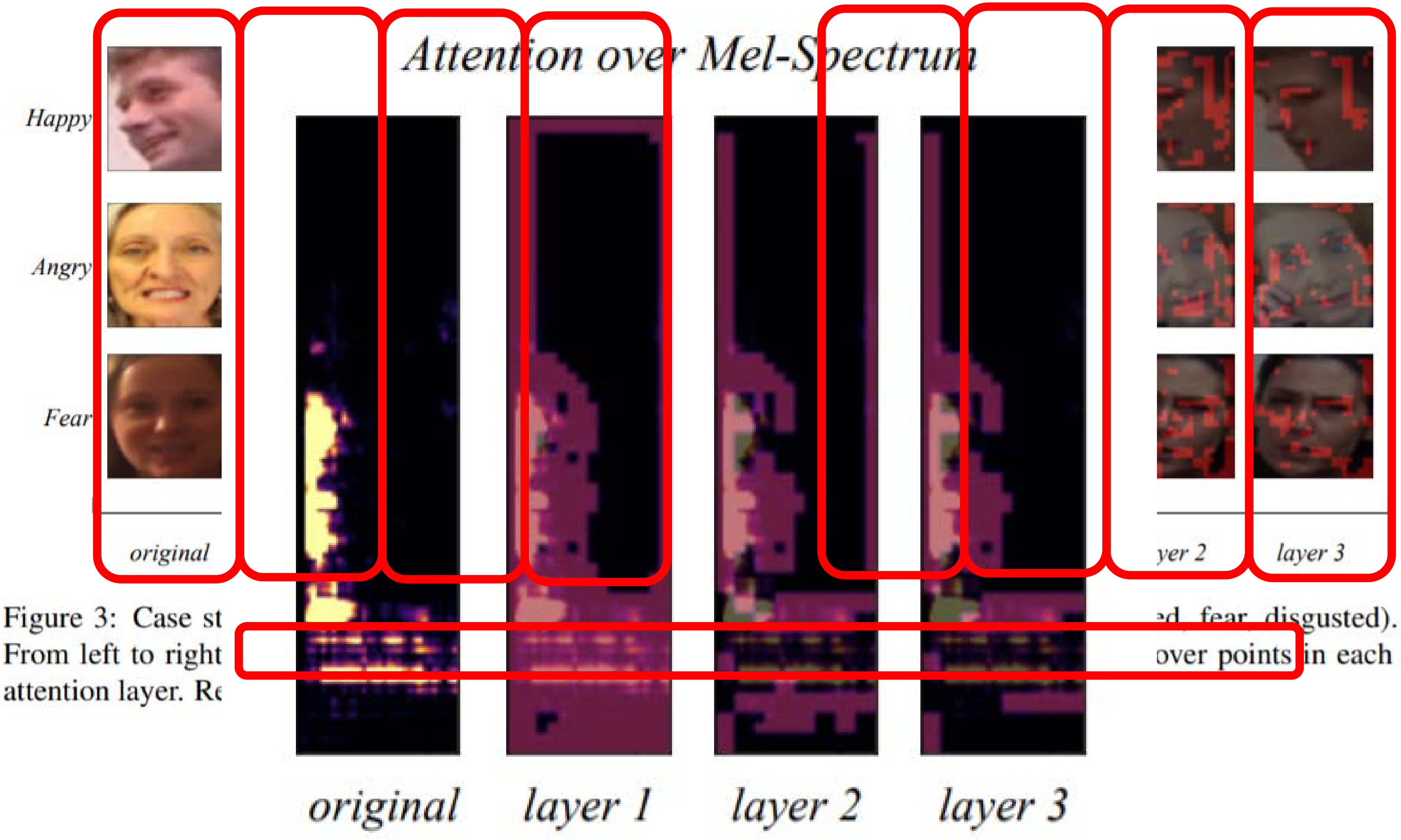
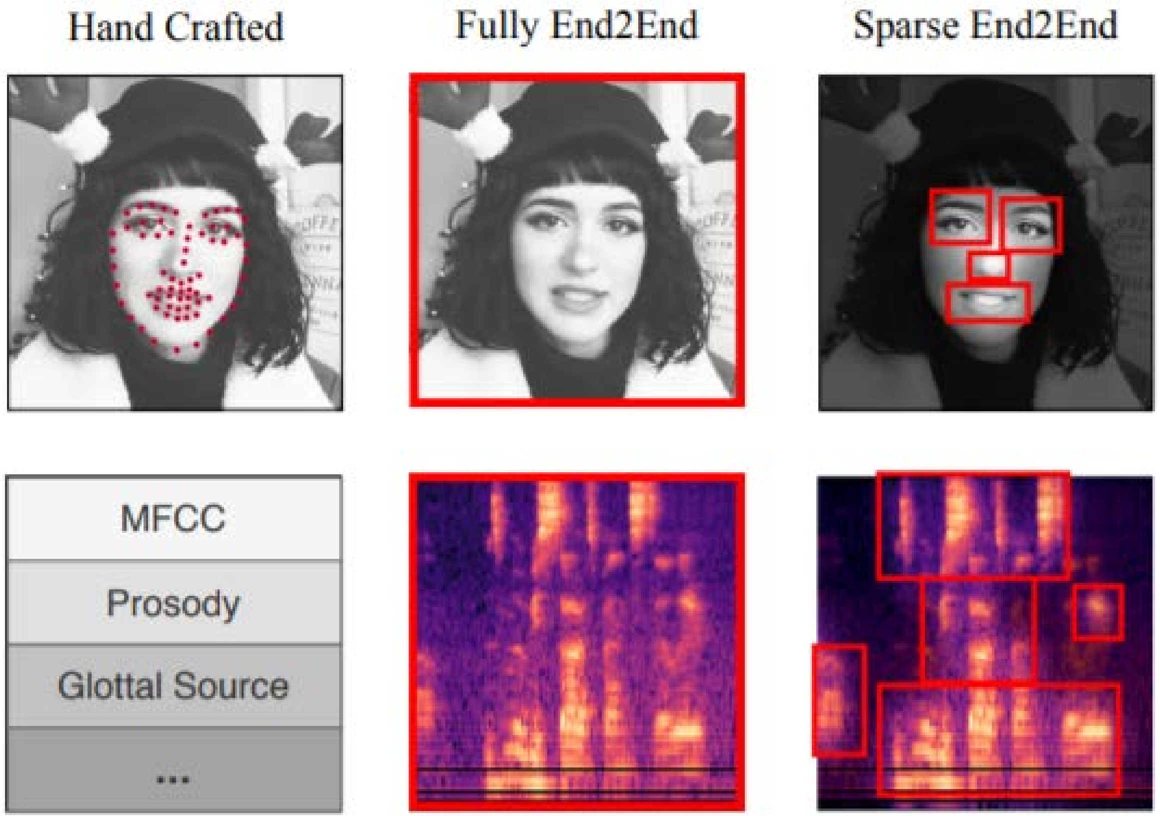


Figure 3: Case st ... ed, fear, disgusted).
From left to right ... over points in each
attention layer. Re

4. Conclusion & Future Work



1) MESM is able to reduce the computational overhead.

2) FE2E model has an advantage in feature learning

   And surpasses the current state of the-art models

3) Visualization of the cross-modal attention maps

   can give an insight to determine modalities.