Information Fusion in Attention Networks Using Adaptive and Multi-Level Factorized Bilinear Pooling for Audio-Visual Emotion Recognition

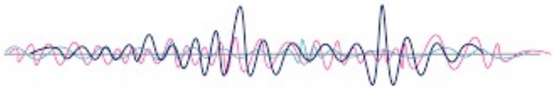
报告人：周恒顺
指导老师：杜俊副教授

Outline

- Introduction
- The Proposed Attention and Fusion Strategy
- Experiment and Result Analyses
- Q&A

Multimodal Emotion Recognition

Video



Audio

Shut up! I don't want to hear anything from you.

Text

Motivation

- Improve the robustness of emotion recognition system in complex scenes.
    - detected face is blurred or occluded
    - audio signal is polluted by noise
- How to fully utilize both audio and visual information is still an open problem.
    - feature-level
    - decision-level fusions
    - model-level fusion

The Proposed Attention and Fusion Strategy

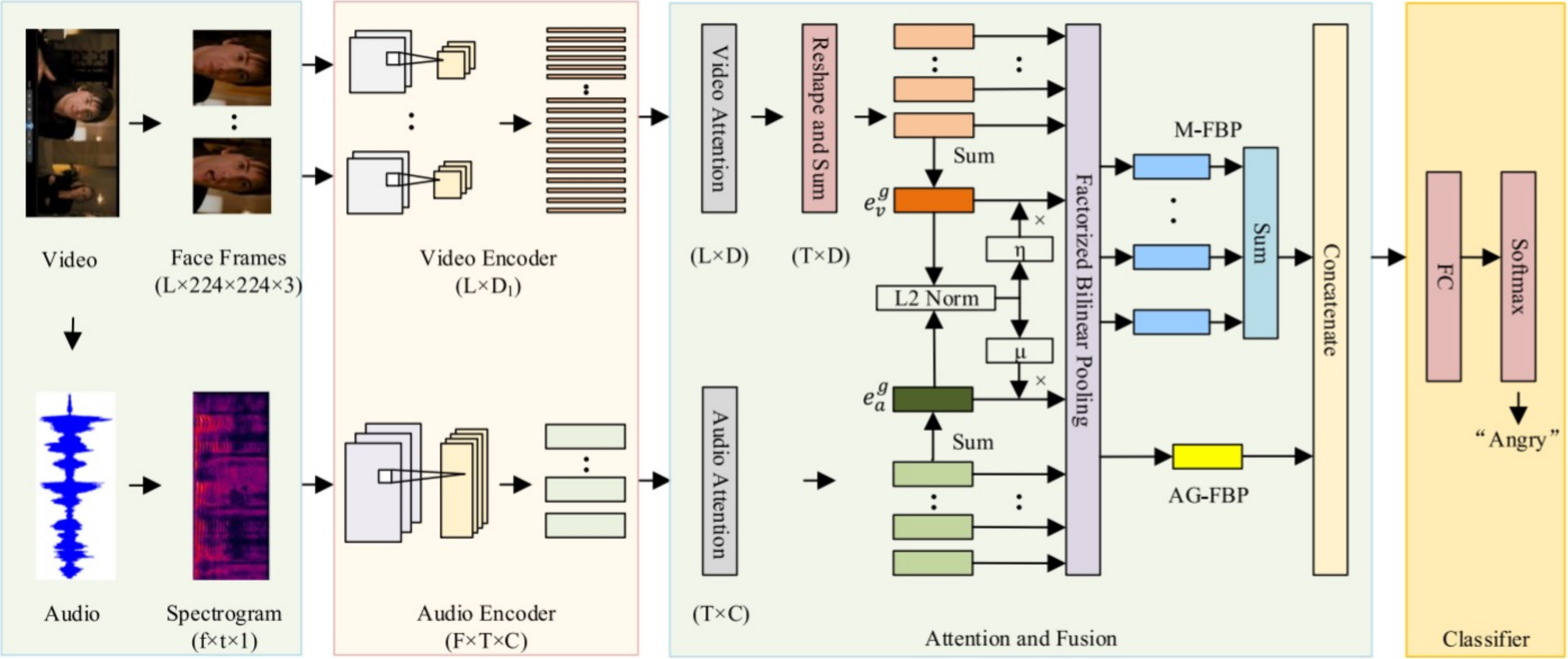


Fig. 1: An overall architecture of the proposed multimodal attention and fusion network based on adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition.

Audio/Video Stream

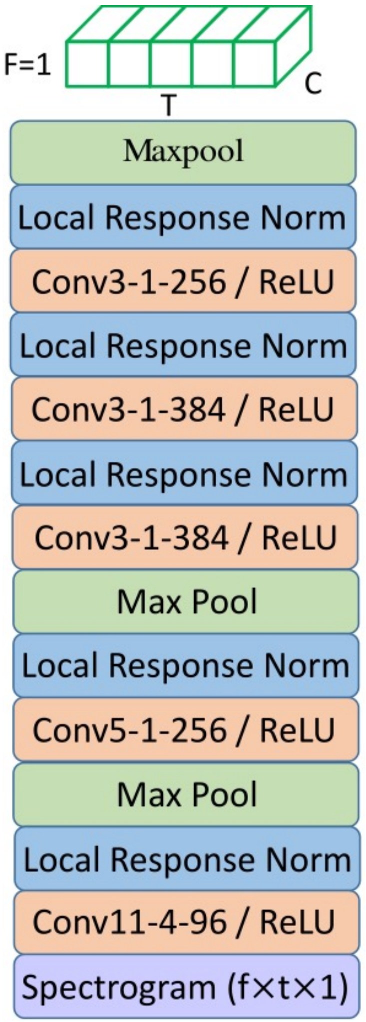


Fig. 2: FCN based audio encoder.

$$\boldsymbol{A} = \{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_T\}, \boldsymbol{a}_i \in \mathbb{R}^C$$

$$\gamma_i^a = \boldsymbol{u}_a^\top \tanh(\boldsymbol{W}_a \boldsymbol{a}_i + \boldsymbol{b}_a)$$

$$\bar{\gamma}_i^a = \frac{\exp(\lambda_a \gamma_i^a)}{\sum_{k=1}^{T} \exp(\lambda_a \gamma_k^a)}$$

$$\boldsymbol{e}_a^i = \bar{\gamma}_i^a \boldsymbol{a}_i$$

$$\boldsymbol{e}_a^g = \sum_{i=1} \boldsymbol{e}_a^i$$

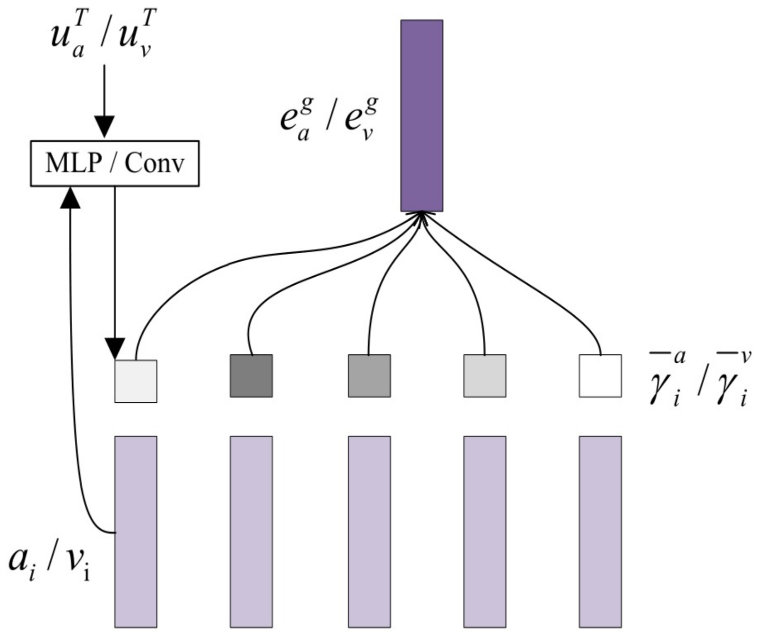


Fig. 3: Structure of audio/video self-attention.

Global Factorized Bilinear Pooling (G-FBP)

For the audio feature vector, $e_a^g \epsilon R^C$, and video feature vector, $e_v^g \epsilon R^D$, the bilinear pooling for the output, $I_j \epsilon R$, is defined as follows:

$$I_j = {e_a^g}^\top \mathbf{\Lambda}_j e_v^g$$

According to [1], the projection matrix Λ_j can be factorized into two low-rank matrices:

$$\begin{aligned} I_j &= {e_a^g}^\top P_j {Q_j}^\top e_v^g \\ &= \sum_{d=1}^{K} {e_a^g}^\top p_j^d {q_j^d}^\top e_v^g \\ &= \mathbb{1}^\top ({P_j}^\top e_a^g \circ {Q_j}^\top e_v^g) \end{aligned}$$

To obtain the output feature vector I below, two 3-D tensors, $P = [P_1, \cdots, P_O] \epsilon R^{C \times K \times O}$ and $Q = [Q_1, \cdots, Q_O] \epsilon R^{D \times K \times O}$, need to be learned. Note P and Q can be reformulated as 2-D matrices, $\tilde{P} \epsilon R^{C \times KO}$ and $\tilde{Q} \epsilon R^{D \times KO}$ respectively.

$$\boldsymbol{I} = \mathrm{SumPooling}(\widetilde{P}^\top e_a^g \circ \widetilde{Q}^\top e_v^g, K_G)$$

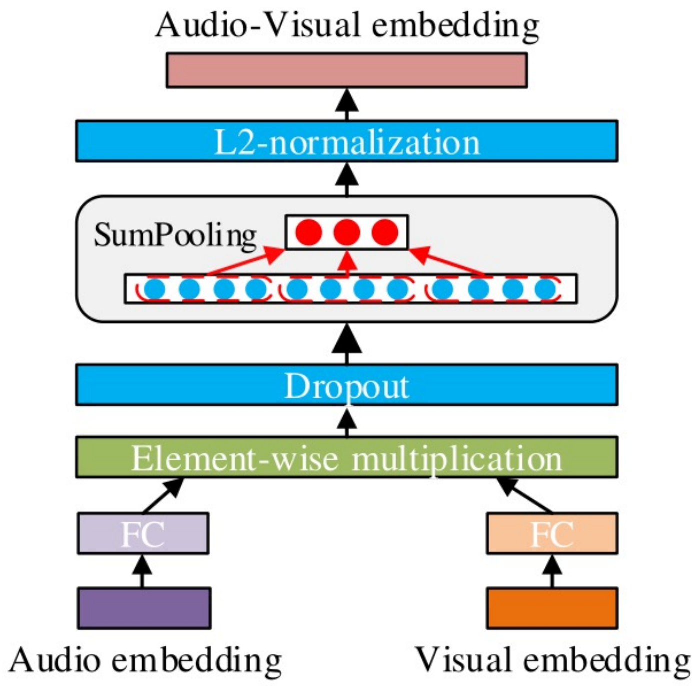


Fig. 4: FBP module

[1] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 1821–1830.

Improved FBP

Adaptive Global Factorized Bilinear Pooling (AG-FBP)

The two coefficients are calculated dynamically by adopting the encoder vectors before audio and video fusion:

$$\mu = \frac{\|\boldsymbol{e}_a^g\|}{\|\boldsymbol{e}_a^g\| + \|\boldsymbol{e}_v^g\|}$$

$$\eta = \frac{\|\boldsymbol{e}_v^g\|}{\|\boldsymbol{e}_a^g\| + \|\boldsymbol{e}_v^g\|}$$

And the new fusion formulation is shown below:

$$I_j^{\mathrm{A}} = (\mu \boldsymbol{e}_a^g)^\top \boldsymbol{\Lambda}_j (\eta \boldsymbol{e}_v^g)$$

Correspondingly, the formulation of G-FBP is modified as:

$$\boldsymbol{I}^{\mathrm{A}} = \mathrm{SumPooling}(\widetilde{\boldsymbol{P}}^\top (\mu \boldsymbol{e}_a^g) \circ \widetilde{\boldsymbol{Q}}^\top (\eta \boldsymbol{e}_v^g), K_G)$$

Improved FBP

Multi-Level factorized bilinear pooling (M-FBP)

To implement M-FBP , the stride of the pooling layer of the audio stream can be modified to adjust the length of intra-trunk audio data, namely $e_a = [e_a^1, \cdots, e_a^H]$. H is the number of intra-trunks and determined by the time lengths of the sample (L) and one intra-trunk (T), and $L = H \times T$.

For the video stream, through the reshape and sum operation, we have $e_v = [e_v^1, \cdots, e_v^H]$.

Finally, we formulate intra-trunk based FBP as follows:

$$\boldsymbol{I}^{\mathrm{M}} = \sum_{h=1}^{H} \mathrm{SumPooling}\left(\widetilde{\boldsymbol{P}}_h^{\top} \boldsymbol{e}_a^h \circ \widetilde{\boldsymbol{Q}}_h^{\top} \boldsymbol{e}_v^h, K_M\right)$$

And both I_A of global-trunk data and I_M of intra-trunk data are concatenated as the fusion vector for the AM-FBP system.

Experiment and Result Analyses

- Data set: IEMOCAP database

4 categories: angry, happy, neutral, sad

Table I: Classification accuracy comparison of different audio network architectures and parameter initializations on IEMOCAP test set.

Systems	Initialization	Accuracy
Att.+BLSTM+FCN [73]	Random	68.10%
CNN+LSTM [5]	Random	68.80%
Fusion_TACN [74]	Random	69.75%
2D-ABFCN [12]	Pre-trained	70.40%
1D-ABFCN (No LRN)	Random	70.79%
1D-ABFCN	Random	71.40%

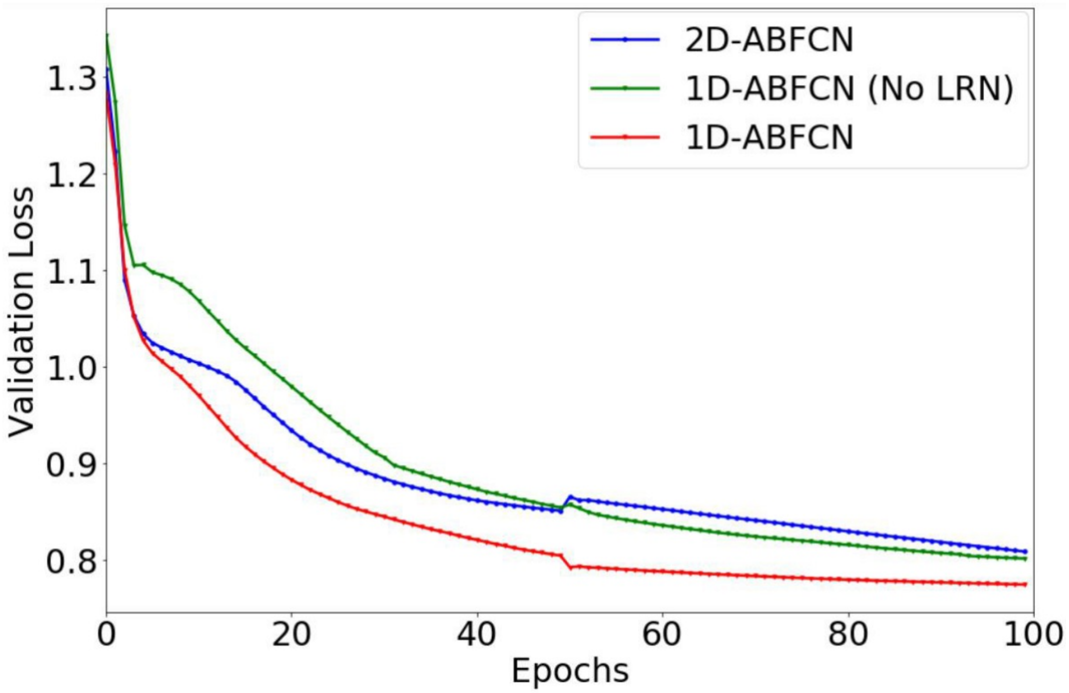


Fig. 5: Comparison of learning curves of different audio emotion recognition systems.

Experiment and Result Analyses

- **Data set:** AFEW database

  7 categories: angry, disgust, fear, surprise, happy, neutral, sad

Table II: Classification accuracy comparison of different systems on AFEW validation set.

Systems	Accuracy
EmotiW2019 baseline [7]	38.81%
Audio system	34.99%
Video system	52.07%
G-FBP	61.10%

Table III: Classification accuracy and p-value of our improved FBP approaches on the AFEW validation set.

Systems	Accuracy	p-value
G-FBP	61.10%	-
AG-FBP	62.40%	0.004
M-FBP	63.18%	0.002
AM-FBP	64.17%	0.001

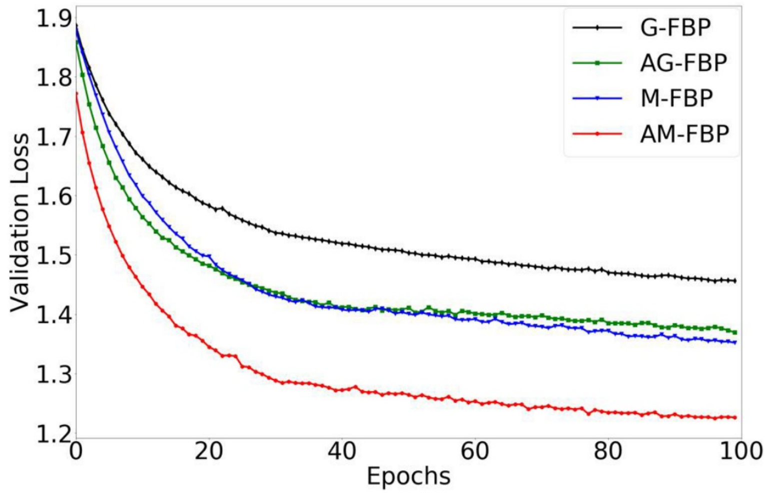


Fig. 6: Learning curves of different FBP systems on the validation set.

Experiment and Result Analyses

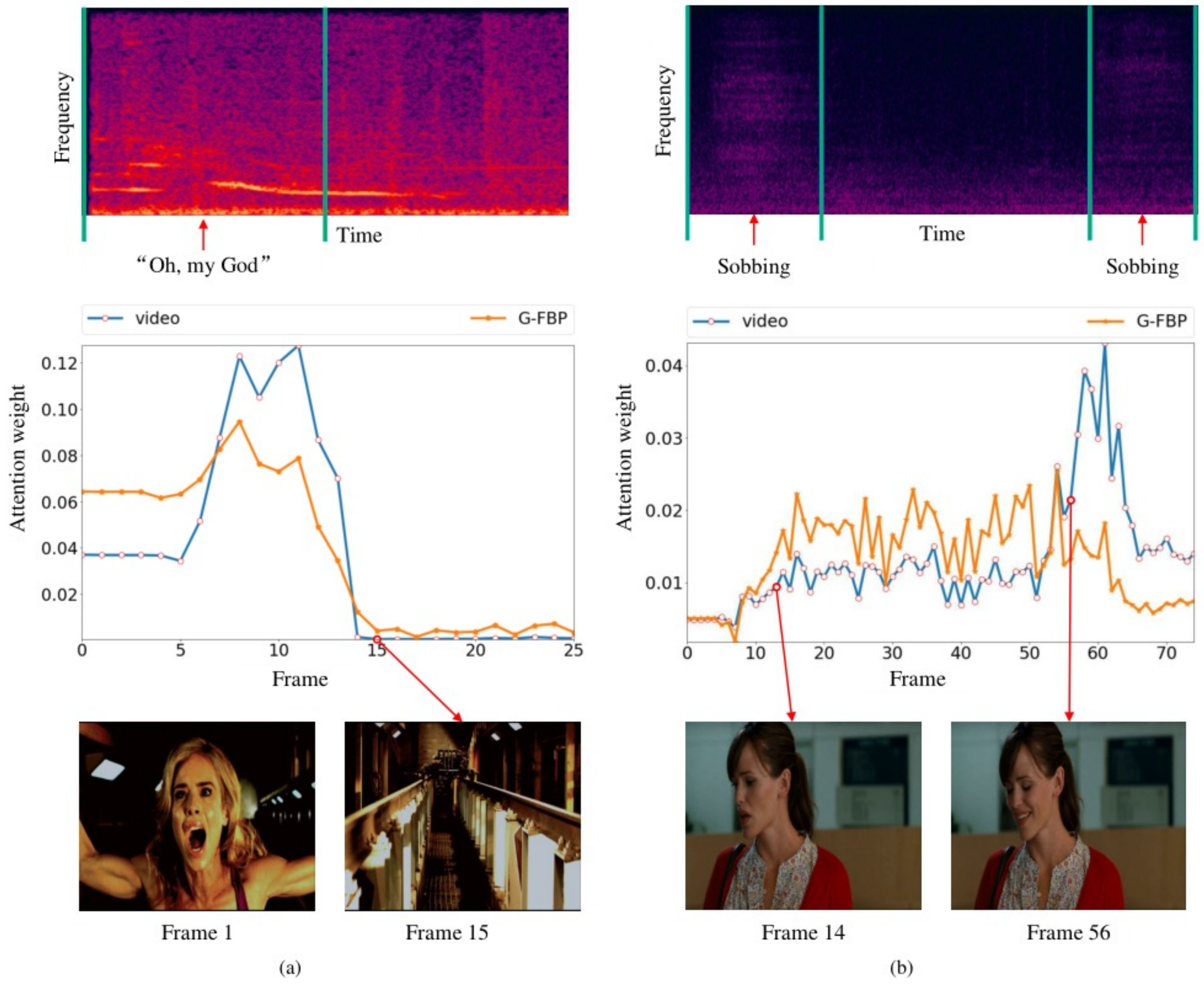


Fig. 7: Attention analysis of two randomly selected examples.

Experiment and Result Analyses

Table V: The overall performance comparison and p-value of different systems on the AFEW test set.

Systems	Single model	Accuracy	p-value
EmotiW2019 baseline [7]	✓	41.07%	-
MAFN [30]	×	58.65%	-
4CNNs+LMED+DL-A+LSTM [29]	×	61.87%	-
4CNNs+BLSTM+Audio [22]	×	62.78%	-
G-FBP	✓	60.64%	-
4G-FBP	×	62.48%	-
AG-FBP	✓	61.26%	0.008
M-FBP	✓	61.87%	0.001
AM-FBP	✓	62.17%	<0.001
2AM-FBP	×	62.79%	<0.001
2AM-FBP+4G-FBP	×	**63.09%**	<0.001

Fig. 10: Confusion matrix on AFEW test set.

Experiment and Result Analyses

Table VI: Classification accuracy comparison and p-value of different systems on IEMOCAP test set.

Systems	Accuracy	p-value
Audio system [78]	63.00%	-
Decision fusion [78]	65.40%	-
Audio system [58]	50.97%	-
Video system [58]	49.39%	-
Encoder concat [58]	67.58%	-
Audio system	71.40%	-
Video system	53.42%	-
Decision fusion	72.54%	-
Encoder concat	73.11%	-
G-FBP	73.98%	0.001
AM-FBP	75.49%	<0.001

Thanks!