ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447231

IMPROVING MULTI-MODAL EMOTION RECOGNITION USING ENTROPY-BASED FUSION AND PRUNING-BASED NETWORK ARCHITECTURE OPTIMIZATION

Haotian Wang[1], *Jun Du*[1*], *Yusheng Dai*[1] ,*Chin-Hui Lee*[2], *Yuling Ren*[3], *Yu Liu*[3]

[1] University of Science and Technology of China, Hefei, Anhui, China
[2] Georgia Institute of Technology, Atlanta, Georgia, USA
[3] China Mobile Online Services Company Limited, China

ABSTRACT

In this study, we aim to improve our recent hierarchical information fusion system for multi-modal emotion recognition challenge (MER 2023) in both efficiency and performance. Specifically, we extract robust acoustic and visual representations from pre-trained models and fuse them together in different structures. Then, an entropy-based fusion approach is proposed to obtain the final prediction of emotion and valence based on multi-label predictions of all different feature fusion structures. Furthermore, to reduce the network redundancy and improve the model generalization in low-resource multi-modal data conditions, we propose a novel approach for optimizing the network structure progressively based on structured pruning and learning-rate rewinding. When tested on the dataset of MER 2023, the optimized network structure with entropy-based fusion yields consistent and significant improvements, outperforming the champion system of the MER-MULTI sub-challenge.

***Index Terms*—** Multi-modal emotion recognition, feature fusion, entropy-based fusion, structured pruning, network architecture optimization

1. INTRODUCTION

Multi-modal emotion recognition (MER) plays an important role in natural human-machine interaction [1], mental health analysis diagnoses [2], intelligent education tutoring [3], etc. Emotions can be calculated by two primary theories: discrete theory and dimensional theory. Discrete theory [4] characterizes emotional states as discrete labels such as “happiness” and “sadness”. Dimensional theory [5] suggests that emotional states exist as points in a continuous space, allowing for the simulation of complex and sustained behaviors. In our research, we applied both theories for emotion recognition.

In human daily lives, emotions are mainly expressed through speech and facial expressions, providing complementary emotive information. Therefore, how to extract emotive acoustic and visual representations and fuse them effectively has become a research hotspot in recent years [6, 7]. Early studies mainly trained the MER systems from scratch [8, 9, 10]. Han *et al.* [8] proposed an approach that maximized the mutual information (MI) among unimodal input pairs. Le *et al.* [9] utilized CNN networks followed by transformer encoders to capture the hidden features from video frames and audio spectrograms and fused them through a transformer-based network. Recently, inspired by the success of pre-trained features such as wav2vec2.0 [11] in other speech-related tasks, some researchers began to investigate their superiority over hand-engineered features and discovered that these deep features captured more robust representations in low-resource multi-modal data conditions [12, 13]. Lian *et al.* [12] conducted a survey on the performance of various speech and image pre-trained models on MER2023 dataset, discovered that acoustic features from HUBERT [14] and visual features from MANet [15] achieved the best results in unimodal emotion recognition, and proposed a fusion framework based on self-attention. Our recent work [13] extended former work by studying the performance difference of deep features from different layers of pre-trained models and proposed a hierarchical information fusion approach. However, current studies on decision-level fusion primarily focus on weighting the decisions from different systems using statistical weights (e.g., linear weighting), while relatively neglecting the variations in samples, which limits the model’s ability to handle certain ambiguous samples. In addition, hand-crafted backend network structures may not obtain optimal performance because redundant connections usually act as noise in evaluation [16], causing confusion issues in the classification of similar emotions.

Pruning is an effective method to remove the redundancy in a network, which can be divided into structured pruning and unstructured pruning [17]. The lottery ticket hypothesis (LTH) [18] revealed the compressible nature of networks. One effective scheme based on LTH is the learning-rate rewinding strategy [19]. Recently, CPLR [20] by integrating channel-level pruning and learning-rate rewinding was proposed and performed well in multi-modal systems. However, the majority of recent researches on pruning usually focuses on how to acquire higher compression ratios, but the potential of optimizing the network structure and improving network performance through structured pruning has rarely been studied.

In this paper, we improve our recent MER system for multi-modal emotion recognition challenge (MER 2023) in both efficiency and performance. First, we extract different levels of deep acoustic features from pre-trained models and separately fuse them with deep visual features. Then we propose an entropy-based fusion approach for combining the multi-label predictions drawn from different fused features to obtain a more reliable decision against confusion issues. Furthermore, to reduce the network redundancy and improve the model generalization in low-resource multi-modal data conditions, a novel approach for optimizing the network structure progressively is proposed based on structured pruning. Our final system outperforms the champion system of MER-MULTI sub-challenge with the highest matrix of 0.7139.

2. METHODS

In this section, we will discuss our proposed multimodal emotion recognition system in two subsections. The architecture of our im-

*corresponding author

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 09,2024 at 14:18:29 UTC from IEEE Xplore. Restrictions apply.

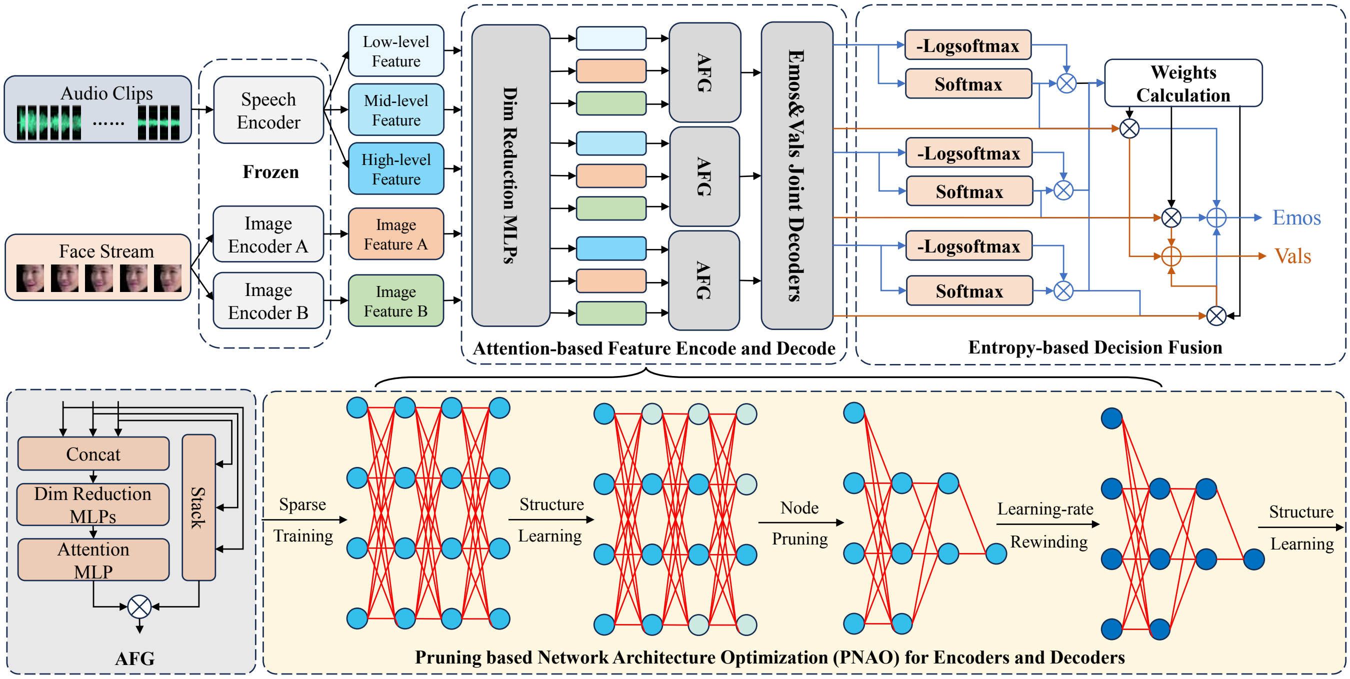


Fig. 1. The architecture of the proposed multi-modal emotion recognition system with pruning-based network architecture optimization (PNAO). AFG represents Attention-guided Feature Gathering in the figure.

proved hierarchical information fusion system with entropy-based decision fusion will be discussed in subsection 2.1. And the principle of the proposed network architecture optimization approach will be illustrated in subsection 2.2. The overall flowchart of our system is shown in Fig. 1, which will be illustrated in following subsections.

2.1. Entropy-based Decision Fusion

In our proposed architecture, robust utterance-level acoustic and visual representations are firstly extracted by pre-trained models from the original feature space. Specifically, following our recent work [13], low-level, mid-level, and high-level acoustic representations are extracted from different layers of HUBERT-large [14]. For the visual part, the pre-trained MANet [15] and ResNet [21] are utilized to obtain complementary visual representations.

Then, as shown in Fig. 1, three distinct acoustic representations are incorporated with visual representations separately in AFG [13, 22] to obtain different levels of acoustic-visual unified representations. Afterward, multi-labels of emotion and valence are predicted with different fused representations in joint decoders [13], which can be formulated as follows, and $i \in 1, 2, 3$ represents different classifiers based on different fused representations:

$$\hat{\boldsymbol{e}}_i = \mathrm{Softmax}(\tilde{\boldsymbol{e}}_i) = \mathrm{Softmax}(\boldsymbol{W}_e \hat{\boldsymbol{h}}_i + \boldsymbol{b}_e) \tag{1}$$

$$\hat{v}_i = \boldsymbol{W}_{vv}[\tilde{v}_{hi}, \tilde{v}_{ei}]^{\mathrm{T}} + b_{vv} \tag{2}$$

where $\hat{\boldsymbol{e}}_i \in \mathbb{R}^C$ (total C emotion categories) and $\hat{v}_i \in \mathbb{R}$ are the predictions of emotion and valence based on single fused representation $\hat{\boldsymbol{h}}_i \in \mathbb{R}^D$. $\boldsymbol{W}_e$, $\boldsymbol{W}_{vv}$, $\boldsymbol{b}_e$, b_{vv} are trainable parameters. $\tilde{v}_{hi} \in \mathbb{R}$ and $\tilde{v}_{ei} \in \mathbb{R}$ are the estimated valence possibilities according to the fused state $\hat{\boldsymbol{h}}_i$ and emotion hidden state $\hat{\boldsymbol{e}}_i$ with trainable parameters $\boldsymbol{W}_{hv}$, b_{hv}, $\boldsymbol{W}_{ev}$ and b_{ev}, calculated as follows:

$$\tilde{v}_{hi} = \boldsymbol{W}_{hv} \hat{\boldsymbol{h}}_i + b_{hv} \tag{3}$$

$$\tilde{v}_{ei} = \mathrm{Tanh}(\boldsymbol{W}_{ev} \tilde{\boldsymbol{e}}_i + b_{ev}) \tag{4}$$

In fact, different levels of fused representations contain various acoustic information [13]. As a result, different emotion classifiers that utilize different levels of fused features can yield varying confidence levels on judgments. Some classifiers may provide a high confidence prediction, while others may provide lower confidence judgments due to the inability to effectively discriminate similar emotions based on the acoustic information they utilize. In order to obtain a more confident judgment, we proposed a confidence-driven approach to obtain a joint prediction based on the predictions of different emotion classifiers, as shown in Fig. 1. We calculate the confidence-level scores of different predictions based on the information entropy of posterior probability predictions on emotion labels, whose principle is as follows:

$$H_i = -\langle \hat{\boldsymbol{e}}_i, \log \hat{\boldsymbol{e}}_i \rangle \tag{5}$$

$$\omega_i = \frac{1}{M-1}\left(1 - \frac{H_i}{\sum_{i=1}^{M} H_i}\right) \tag{6}$$

where H_i is the information entropy of each emotion prediction and ω_i is the confidence-level score based on entropy H_i. Higher entropy, lower confidence-level score. M represents the amount of predictions ($M = 3$ in our framework). Then we get the joint decision by weighting the posterior probabilities of emotion and valence predictions based on their confidence-level scores, as follows:

$$\begin{bmatrix} \hat{\boldsymbol{e}} \\ \hat{v} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{e}}_1 & \ldots & \hat{\boldsymbol{e}}_M \\ \hat{v}_1 & \ldots & \hat{v}_M \end{bmatrix} \cdot \begin{bmatrix} \omega_1 & \ldots & \omega_M \end{bmatrix}^{\mathrm{T}} \tag{7}$$

where $\hat{\boldsymbol{e}} \in \mathbb{R}^C$, $\hat{v} \in \mathbb{R}$ is the joint prediction for emotion and valence. Higher weights are assigned to more confident predictions when fusion. The experiments in subsection 3.2 show that we can alleviate the confusion situation of similar emotions and improve the system performance by trusting the most confident predictions.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 09,2024 at 14:18:29 UTC from IEEE Xplore. Restrictions apply.

2.2. Pruning-based Network Architecture Optimization (PNAO)

Compared to speech emotion recognition data, the multi-modal emotion data are more low-resource. The redundancy in our initially designed multi-modal emotion recognition system may lead to poor performance. To reduce the network redundancy and improve the model generalization in low-resource multi-modal data conditions, we proposed a novel approach for optimizing the fine-grained network structure progressively. The details are shown in Algorithm 1.

Algorithm 1 PNAO algorithm

---

1 : Pre-train the initial network parameter matrices to the early-stop point $\mathbf{\Theta}^0$ by using sparse-training.
2 : Set network architecture optimization rate (NAOR) k.
3 : Learn the mask $\boldsymbol{m}$ based on L_1 norm of row dimension of all parameter matrices $\mathbf{\Theta}^0$ with NAOR k.
4 : Prune the nodes of network using the mask, obtain the reinit network parameter matrices $\mathbf{\Theta}^0 \odot \boldsymbol{m}$.
5 : Use learning rate rewinding strategy to retrain the network to obtain the fine-tuned parameter matrices $\mathbf{\Theta}^1$.
6 : Repeat **2** to **5** for N rounds to obtain the best-performing compact network.

Firstly, we train the initial network to the early-stop point [18] with sparse-training. During training, we use the cross-entropy (CE) loss as the emotion classification loss, denoted as $\mathcal{L}_e$, and the mean squared error (MSE) loss is adopted for valence prediction, denoted as $\mathcal{L}_v$. Additionally, we introduce uncertainty loss weighting [23] to $\mathcal{L}_e$ and $\mathcal{L}_v$ for better performance in the multi-task learning process, denoted as AWL. Without loss of generality, the total loss function at a certain round is as follows, where $\mathbf{\Theta} = \{\mathbf{\Theta}_1, ..., \mathbf{\Theta}_L\}$ (total L layers in network) represents all parameter matrices of this round:

$$\mathcal{L}_{ev} = \mathrm{AWL}(\mathcal{L}_e, \mathcal{L}_v) + \alpha \cdot \|\mathbf{\Theta}\|_p \tag{8}$$

$$\mathrm{AWL}(\mathcal{L}_e, \mathcal{L}_v) = \frac{1}{{\delta_1}^2}\mathcal{L}_e + \frac{1}{2{\delta_2}^2}\mathcal{L}_v + \log(1+\delta_1) + \log(1+\delta_2) \tag{9}$$

The sparse term based on the p-norm (p=1 in the equation) is added to the loss function, and sparse training has been proven to be effective for dynamic pruning networks [24]. In the training process, the unimportant weights will become smaller and smaller, making the less important nodes more and more discriminative.

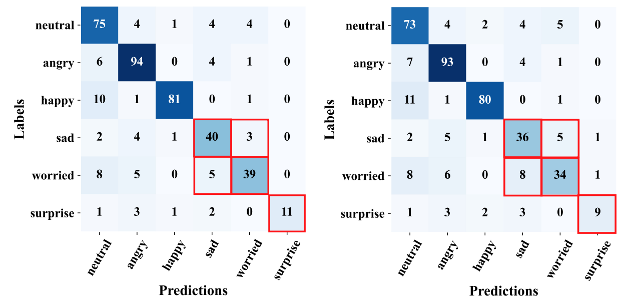
Then, we learn the mask based on network architecture, we adopt the L_1 norm of column dimension of parameter matrices as indicating factors to node importance, as follows:

$$\boldsymbol{\gamma}(l,j) = \|\mathbf{\Theta}_l(j,:)\|_1 = \frac{1}{N_l}\sum\nolimits_{n=1}^{N_l} |\mathbf{\Theta}_l(j,n)| \tag{10}$$

where l denotes the index of the current layer and j and r denote the indexes of nodes and columns (total N_l columns) in the l-th layer. $\boldsymbol{\gamma}$ represents the importance matrix of total nodes. The global mask matrix $\boldsymbol{m}$ is generated based the network architecture optimization rate (NAOR) k and importance matrix $\boldsymbol{\gamma}$, computed as follows:

$$\boldsymbol{m}(l,j) = \mathrm{U}(\boldsymbol{\gamma}(l,j) - k \cdot \max(\boldsymbol{\gamma})) \tag{11}$$

$\mathrm{U}(\cdot)$ is the unit step function. Next, the mask $\boldsymbol{m}$ is applied to matrices $\mathbf{\Theta}$ and the zero-weighted nodes are removed. The more compact network is then fine-tuned using the learning-rate rewinding strategy [19]. The following steps are repeated several times until the optimal network has been found. By adopting a small NAOR, we can prune off some redundant nodes while retaining the important nodes, so the network architecture will be optimized step by step.



(a) Entropy-based fusion strategy. (b) Attention-based fusion strategy.

Fig. 2. Performance comparison of entropy-based fusion strategy and former attention-based strategy.

3. EXPERIMENTS AND RESULTS ANALYSIS

In this section, several experiments are conducted to validate the effectiveness of the proposed methods. Similar to our previous work [13], the outputs of the $18, 19$ and 20-th layers of HUBERT-large [14] are adopted as acoustic representations and the outputs of MANet [15] and ResNet [21] are adopted as visual representations in all systems for fair comparisons in all the following experiments.

3.1. Dataset and Metric

In this research, we conduct experiments on MER 2023 dataset [12]. The dataset consists of 3373 labeled single-speaker video segments used as the training dataset. There are 411 and 412 unlabeled video segments for the test set in MER-MULTI sub-challenge. Same with the baseline [12], the combined metric (Com) of emotion classification (Dis) and valence regression (Dim) is chosen to evaluate the overall performance of discrete and dimensional emotions.

3.2. Effectiveness of Entropy-based Fusion

To evaluate the effectiveness of the proposed entropy-based decision fusion strategy on a hierarchical information fusion system, we plot the 6-class emotion classification confusion matrices of our improved MER system and the well-performing attention-based fusion strategy of our former MER system that ranks third on MER-MULTI [13]. The two systems share the same feature fusion architecture but different decision fusion strategies. As shown in Fig. 2, the results suggest that the entropy-based fusion strategy performs better than the former attention-based fusion strategy, obtaining lower confusion between emotions. It is worth noticing that the entropy strategy remarkably improved the classification of easily confused emotion categories, obtaining more correctly classified samples of sad, worried and surprised emotions. Benefiting from the proposed entropy-based decision fusion strategy, the high confidence prediction can attribute more to the final prediction when facing confusing samples, while enhancing the classification ability of our MER model against confusion situations on emotions.

3.3. Effectiveness of PNAO

In this section, we further conduct PNAO on the entropy-based MER system tested in subsection 3.2 to progressively find the optimal architecture. We set the network architecture optimization rate to 0.05 and proceed with 10 rounds of optimization. As shown in Fig. 3, the results show that continuous improvements in performance have

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 09,2024 at 14:18:29 UTC from IEEE Xplore. Restrictions apply.

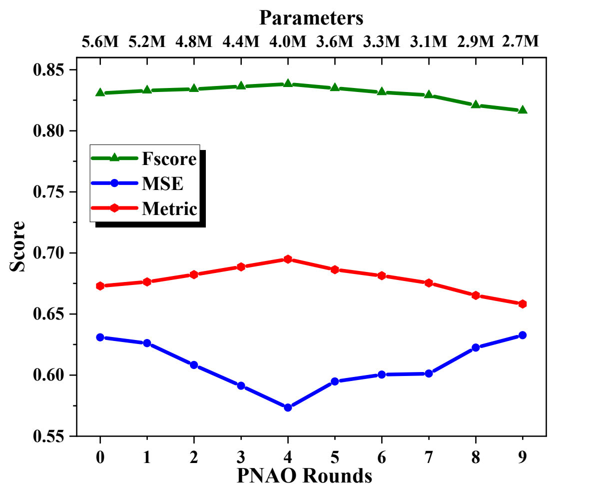


Fig. 3. The performance curve of PNAO on MER system with entropy-based fusion over rounds.

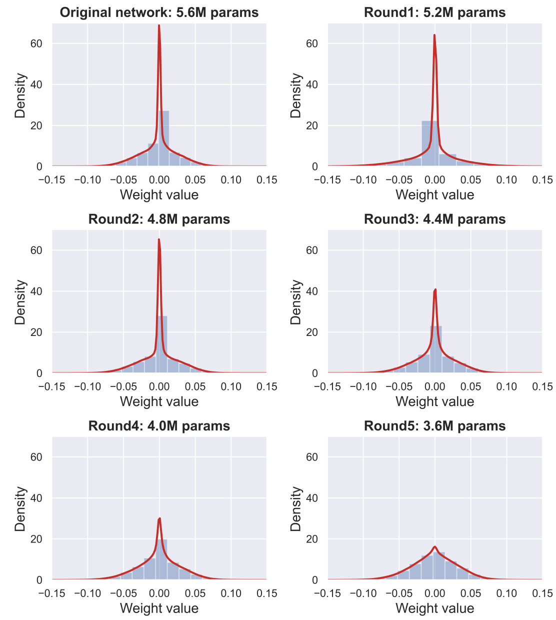


Fig. 4. The weight contribution comparison over rounds of PNAO.

Table 1. Performance comparison of MER systems. EnF denotes entropy-based fusion.

Model	Train&Vals	MER-MULTI		
	Com (↑)	Dis (↑)	Dim (↓)	Com (↑)
sense-dl-lab [25]	-	-	-	**0.7005**
AIPL-SEU [26]	-	-	-	0.6860
USTC-qw [13]	0.6402	0.8328	0.5930	0.6846
Baseline	0.6267	0.8287	0.6033	0.6779
Ours(EnF)	0.6375	0.8350	0.6258	0.6786
Ours(PNAO)	0.6481	0.8301	0.5765	0.6860
Ours(EnF+PNAO)	0.6579	0.8383	0.5734	0.6949
Ours(Final)	**0.6762**	**0.8530**	**0.5563**	**0.7139**

been achieved in the first few rounds of PNAO with decreasing parameters. The Fscore of emotions is increasing and the MSE loss is decreasing progressively. This suggests that the redundant connections in the original network have been a limitation to system performance, and PNAO successfully optimizes the network structure and improves the model generalization by pruning off these connections step by step. The system achieves a Metric score of 0.6949, demonstrating an improvement of 2.2% with a reduction of 28.3% in parameters compared to the original system.

It is also worth noticing that the Metric will decrease after 5 rounds of PNAO, which might be due to the important information lost with the parameters over-pruning. To further investigate this phenomenon, we analyze the dynamic progression of the weight value distribution during rounds, as visualized in Fig. 4. It is observed that the sharp distribution of weight values is progressively pruned to become smoother over rounds. During the first few rounds, redundant weights near zero are pruned and useful weights are activated at the same time, which is a possible explanation for the performance improvement in model generalization. However, weights tend to be averaged after a few rounds along with the reduction of redundant nodes. It is difficult to distinguish insignificant connections and some essential connections may be incorrectly pruned off in further optimization, leading to a decline in system performance.

3.4. Overall Comparison

Finally, we give an overall comparison of our proposed system with other state-of-the-art systems on the MER-MULTI leaderboard [12]. Table 1 presents the top three fusion systems on the MER-MULTI sub-challenge, which are sense-dl-lab [25], AIPL-SEU [26], and USTC-qw [13]. The baseline system is the best single system from USTC-qw [13]. The results indicate that by combining the proposed entropy-based fusion strategy and the PNAO strategy, the single system performance obtains a score gain of 1-2 percent points, indicating the effectiveness of the two proposed techniques. Finally, by combining the decisions of the single systems optimized by PNAO (Ours(PNAO) and Ours(EnF+PNAO) in Table 1) at the decision level using linear weighting, our fusion system (Ours(Final)) achieved the highest metric of 0.7139, which is an improvement of 1.34% compared to the champion system on MER-MULTI.

4. CONCLUSIONS

In this study, we improve our recent hierarchical information fusion system in both efficiency and performance. Firstly, feature fusion structures are designed based on different levels of deep features extracted from pre-trained models. Then we propose an entropy-based decision fusion approach for better integrating the multi-label predictions of different feature fusion structures, obtaining highly-confident decisions of emotion and valence against confusing issues. Furthermore, we proposed a novel approach named PNAO to optimize the structure of the proposed MER system progressively in low-resource training conditions. When tested on the MER 2023 dataset, the final optimized network with entropy-based decision fusion achieves state-of-the-art performance on MER-MULTI.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant 62171427.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 09,2024 at 14:18:29 UTC from IEEE Xplore. Restrictions apply.

6. REFERENCES

[1] Fatemeh Noroozi, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and Gholamreza Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60–75, 2019.

[2] Anoop K, Deepak P, and Lajish V L, "Emotion cognizance improves health fake news identification," in *Proceedings of the 24th Symposium on International Database Engineering & Applications*, New York, NY, USA, 2020, IDEAS '20, Association for Computing Machinery.

[3] Mona Hafez Mahmoud, "A survey of some interdisciplinary methods and tools to measure learners' emotions in intelligent tutoring systems," in *2019 6th International Conference on Advanced Control Circuits and Systems (ACCS) and 2019 5th International Conference on New Paradigms in Electronics & information Technology (PEIT)*, 2019, pp. 1–6.

[4] Jessica L Tracy and Daniel Randles, "Four models of basic emotions: A review of ekman and cordaro, izard, levenson, and panksepp and watt," *Emotion review*, vol. 3, no. 4, pp. 397–405, 2011.

[5] Tanmayee Joshi, Sarath Sivaprasad, and Niranjan Pedanekar, "Partners in crime: Utilizing arousal-valence relationship for continuous prediction of valence in movies.," in *AffCon@ AAAI*, 2019, pp. 28–38.

[6] Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer, "Emotion recognition from multiple modalities: Fundamentals and methodologies," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 59–73, 2021.

[7] Darshana Priyasad, Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes, "Attention driven fusion for multi-modal emotion recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3227–3231.

[8] Wei Han, Hui Chen, and Soujanya Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," 2021.

[9] Hoai-Duy Le, Guee-Sang Lee, Soo-Hyung Kim, Seungwon Kim, and Hyung-Jeong Yang, "Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning," *IEEE Access*, vol. 11, pp. 14742–14751, 2023.

[10] Hengshun Zhou, Jun Du, Yuanyuan Zhang, Qing Wang, Qing-Feng Liu, and Chin-Hui Lee, "Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 2617–2629, jul 2021.

[11] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[12] Zheng Lian, Haiyang Sun, Licai Sun, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, et al., "Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning," 2023.

[13] Haotian Wang, Yuxuan Xi, Hang Chen, Jun Du, Yan Song, Qing Wang, Hengshun Zhou, Chenxi Wang, Jiefeng Ma, Pengfei Hu, Ya Jiang, Shi Cheng, Jie Zhang, and Yuzhe Weng, "Hierarchical audio-visual information fusion with multi-label joint decoding for mer 2023," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, MM '23.

[14] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[15] Jingyun Liang, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, "Mutual affine network for spatially variant kernel estimation in blind image super-resolution," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 4076–4085.

[16] Haotian Wang, Jun Du, Hengshun Zhou, Chin-Hui Lee, Yuling Ren, and Jiangjiang Zhao, "A Multiple-Teacher Pruning Based Self-Distillation (MT-PSD) Approach to Model Compression for Audio-Visual Wake Word Spotting," in *Proc. INTERSPEECH 2023*, 2023, pp. 2678–2682.

[17] Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie, "Model compression and hardware acceleration for neural networks: A comprehensive survey," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 485–532, 2020.

[18] Jonathan Frankle and Michael Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *ICLR*, 2019.

[19] Alex Renda, Jonathan Frankle, and Michael Carbin, "Comparing rewinding and fine-tuning in neural network pruning," in *ICLR*, 2020.

[20] Haotian Wang, Jun Du, Hengshun Zhou, Heng Lu, and Yuhang Cao, "A novel approach to structured pruning of neural network for designing compact audio-visual wake word spotting system," in *APSIPA ASC*, 2022, pp. 820–826.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] Zheng Lian, Jianhua Tao, Bin Liu, and Jian Huang, "Conversational Emotion Analysis via Attention Mechanisms," in *Proc. Interspeech 2019*, 2019, pp. 1936–1940.

[23] Roberto Cipolla, Yarin Gal, and Alex Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.

[24] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang, "Learning efficient convolutional networks through network slimming," in *ICCV*, 2017, pp. 2755–2763.

[25] Daoming Zong, Chaoyue Ding, Baoxiang Li, Dinghao Zhou, Jiakui Li, Ken Zheng, and Qunyan Zhou, "Building robust multimodal sentiment recognition via a simple yet effective multimodal transformer," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, MM '23.

[26] Sunan Li, Hailun lian, Cheng Lu, Yan Zhao, Chuangao Tang, Yuan Zong, and Wenming Zheng, "Multimodal emotion recognition in noisy environment based on progressive label revision," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, MM '23.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 09,2024 at 14:18:29 UTC from IEEE Xplore. Restrictions apply.