SPEECH EMOTION RECOGNITION WITH GLOBAL-AWARE FUSION ON MULTI-SCALE FEATURE REPRESENTATION

Wenjing Zhu, Xiang Li*

Du Xiaoman, Beijing, China
{zhuwenjing01,lixiang01}@duxiaoman.com

ABSTRACT

Speech Emotion Recognition (SER) is a fundamental task to predict the emotion label from speech data. Recent works mostly focus on using convolutional neural networks (CNNs) to learn local attention map on fixed-scale feature representation by viewing time-varied spectral features as images. However, rich emotional feature at different scales and important global information are not able to be well captured due to the limits of existing CNNs for SER. In this paper, we propose a novel **GL**obal-**A**ware **M**ulti-scale (GLAM) neural network[1] to learn multi-scale feature representation with global-aware fusion module to attend emotional information. Specifically, GLAM iteratively utilizes multiple convolutional kernels with different scales to learn multiple feature representation. Then, instead of using attention-based methods, a simple but effective global-aware fusion module is applied to grab most important emotional information globally. Experiments on the benchmark corpus IEMOCAP over four emotions demonstrates the superiority of our proposed model with 2.5% to 4.5% improvements on four common metrics compared to previous state-of-the-art approaches.

Index Terms— Speech Emotion Recognition, Attention Mechanism, Multi-scale Features, Feature Fusion

1. INTRODUCTION

Speech is the key medium of communication between people and plays an important role in everyone's daily life. Speech Emotion Recognition (SER) aims to predict the emotion reflected in speech. SER is a fundamental task for various intelligent applications. For example, SER brings better user experience for intelligent robots by better understanding user's intents and states. Thus, interests from research communities have been increased for SER task [1].

Recently, inspired by the uplifting progress in computer vision, existing studies [2, 3, 4, 5, 6, 7, 8, 9] have achieved great improvements on SER by viewing spectral features as images. The standard convolutional neural network (CNN)

architecture mainly consists of feature representation and attention map. Generally, feature representation usually learns fixed-scale features for capturing all informative features. Besides, attention map is utilized to attend most emotional information among local fixed-scale features [6, 7, 8, 9]. However, emotions are characterized by articulation variations with different scales in time-varied spectral feature.

Due to the limits of existing CNNs in SER, we propose a novel **GL**obal-**A**ware **M**ulti-scale (GLAM) neural network. In our work, we utilize multi-scale convolutional blocks to extract features at varying scales due to different temporal spans and intonation strength in emotions. Furthermore, a global-aware fusion module is introduced to select important information reflecting emotional patterns in different scales. Experiments show both multi-scale feature representation and global-aware fusion can bring great improvements compared to the state-of-the-art CNN model[9].

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to apply multi-scale feature representation to SER. Besides, this representation also enhances existing CNNs greatly.

- We introduce a new global-aware fusion module to grab emotional information across multiple scales. Our study displays the great potential of information fusion among different granularities.

- Experiments demonstrate the superiority of our proposed model on four common metrics for SER. We achieve 2.5% to 4.5% improvements on the IEMOCAP corpus compared to state-of-the-art methods.

The rest of this paper is organized as follows. We review related methods in Section 2. Section 3 presents the proposed GLAM neural network. Finally, Section 4 and Section 5 are the experiment results and conclusion.

2. RELATED WORK

Typical works on SER focus on feature representation of the emotion information in speech. The first deep learning

---

*Corresponding author.
[1]The code is available at https://github.com/lixiangucas01/GLAM

method was utilized to learn feature representation for predicting emotion label of each feature segment [10]. Then, by viewing feature segments as images, convolutional networks show the great effects on feature representation for SER [2, 4, 6, 8, 9, 11, 12]. Finally, various methods such as extreme learning machine (ELM) [10], capsule networks [12], and LSTM [2, 11] have been applied to fuse all feature representations from all feature segments. However, most existing CNNs focus on fixed-scale local feature and ignore the effects of varied scales, which are fully considered in our proposed model.

Recently, attention-based models have made significant progress on SER [6, 7, 8, 9, 13]. For example, multi-head attention maps are learned by convolutional operations to select important information according to surrounding information [8]. Moreover, area attention is further introduced to compute the importance from different ranges of convolutions [9]. Though local information is useful for describing subtle information, the lack of perceiving global information limits the advance of existing attention-based methods. Hence, we propose a global-aware fusion module to compute global attention map and fuse emotional information across different scales.

Fig. 1. Overview of the model architecture.

3. METHODOLOGY

In this section, we present our GLAM model in detail. As shown in Fig. 1, GLAM contains two main modules: multi-scale module and global-aware fusion module. We first introduce these two modules and then combine a *mixup* method to enhance the generalization of GLAM.

3.1. Multi-scale Feature Representation

We propose a multi-scale block to parallelize multi-scale convolutional layers on the same level. The multi-scale block, which receive different sized receptive fields, can capture multi-scale feature representation. The structure is shown in right down panel of Fig. 1. We keep the channel size unchanged. For better capturing spatial and temporal features, we implement kernel size of 1×3 for spatial convolution and kernel size of 3×1 for temporal convolution with batch normalization put after on. The out-channels both are 16. For the first layer, we concatenate spatial and temporal convolved outputs alongside spatial dimension. For the rest, the outputs are concatenated alongside channel dimension. In the last convolutional layer with batch normalization followed, we use a large kernel size of 5×5. Same paddings are applied to all of convolutional layers.

3.2. Global-Aware Fusion

As the multi-scale architecture extracts local features, we consider correlating features and enhance feature communica-
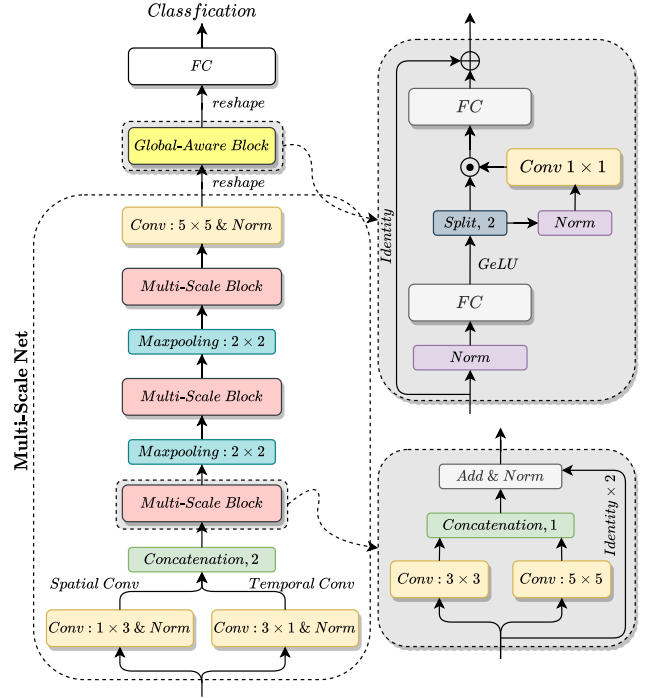
tions globally. We implement a gMLP block [14], named as an global-aware block in this paper, shown in the right up panel of Fig. 1, to enhance cross-scale feature communication. For a feature map X with a size of $C \times d_f$, where C is the channel size and d_f represents size of flattened feature map. The output dimension of first and last fully connected layer are $4d_f$ and d_f respectively. After a projection of GeLU, the output is split along feature dimension. And the convolutional layer keeps channel size unchanged. The operations of multiplication enhance feature mixing across channel dimension. The output of global-aware fusion module is then reshaped to feed into fully connected (FC) network for classification.

3.3. Data Augment

To improve generalization capability of the system, we use a *mixup* method [15] for training, which combines pairs of examples and their labels. *Mixup* mixes two pair training examples (x_i, y_i) and (x_j, y_j) as a new constructed example

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j \tag{1}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j \tag{2}$$

where (x_i, y_i) and (x_j, y_j) are randomly drawn from training data and $\lambda \sim Beta(\alpha, \alpha)$ with $\alpha \in (0, \infty)$. *Mixup* regularizes the neural network to favor simple linear behavior in-between training examples. This method effectively smoothens discrete data space into continuous space.

Table 1. Comparison of evaluation metrics on three types of datasets.

Dataset	Model	WA	UA	macro F1	micro F1
Improvisation	APCNN	70.73±2.57	67.96±2.66	68.26±3.17	70.20±3.00
	MHCNN	76.09±2.01	73.87±2.31	74.27±2.25	75.91±2.05
	AACNN	78.47±2.42	76.68±3.29	76.69±3.27	78.29±2.58
	GLAM	**81.18±1.47**	**79.25±1.88**	**79.87±1.64**	**80.99±1.50**
Script	APCNN	55.95±2.93	54.86±2.74	48.75±3.71	51.15±3.87
	MHCNN	64.57±2.12	63.42±2.14	61.46±2.50	63.36±2.56
	AACNN	67.20±4.78	65.94±6.19	64.82±7.63	66.55±6.99
	GLAM	**71.44±2.05**	**70.39±2.13**	**69.56±2.13**	**70.91±2.12**
Full	APCNN	62.53±1.43	62.80±1.53	61.89±1.52	62.12±1.50
	MHCNN	69.80±1.69	70.09±1.74	69.65±1.73	69.68±1.71
	AACNN	70.94±5.08	71.04±4.84	70.59±6.34	70.71±6.44
	GLAM	**73.70±1.25**	**73.90±1.31**	**73.51±1.29**	**73.60±1.27**

4. EXPERIMENTS

In this section, we first introduce the experimental settings. Then, we compare our proposed model with state-of-the-art methods. Finally, we conduct ablation study on two modules.

4.1. Experimental Settings

Datasets. We conduct our experiments on the benchmark corpus IEMOCAP [16]. IEMOCAP contains 12 hours of emotional speech performed by 10 actors. We select four types of emotions *angry*, *happy*, *sad*, and *neutral* as same as previous studies [6, 7, 8, 9, 13] for balancing the category distribution. In agreement with previous works, we label *excited* as *happy* category due to similarity of two categories. According to whether the actors perform a fixed script, we have three types of datasets for experiments: Improvisation, Script, and Full.

Compared Methods and Evaluation Metrics. To study our proposed model deeply, we choose the best model of Area Attention CNN (AACNN) [9] with recent work Multi-Head Attention CNN (MHCNN) [8] and Attention Pooling CNN (APCNN) [6] for comparison. Moreover, to comprehensively understand prediction performance, we utilize four common metrics Weighted Accuracy (WA), Unweighted Accuracy (UA), Micro F1 score, and Macro F1 score. The distinction between WA and UA (micro and macro F1 score) is that the former takes label imbalance into account, while the latter does not. We report the mean value and standard deviation of above metrics in experiments. In addition, we will also show confusion matrix for concrete analysis.

Implementation Details. Since there is no unified way to split dataset, we randomly divide dataset into 80% of data for training and the rest for test with 100 times to ensure the reliability of experiment results. We use mel-frequency cepstral coefficients (MFCCs) [17, 18] as the feature input. Each utterance is divided into 2-second segments with 1.6 second over-

lap between segments. We average the prediction results of all segments in same utterance as the prediction result. Cross-entropy criterion is used as an objective function. Adam optimizer is adopted with weight decay rate as 10^{-6}. The learning rate is initialized to 10^{-4} and exponentially decayed with multiplicative factor 0.95 until the value reaches 10^{-6}. The models are all trained for 50 epochs and the batch size is set to 32. In our proposed model, we evaluate the performance of different mixup parameter α and set it to 0.5 for comparison.
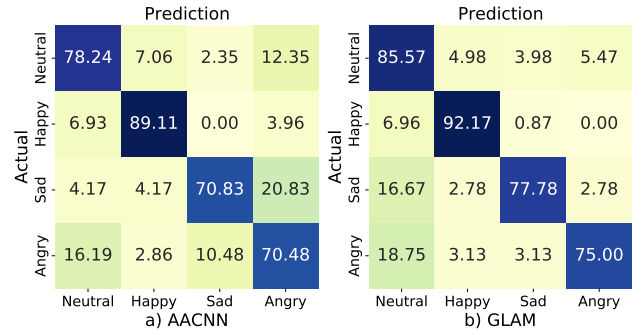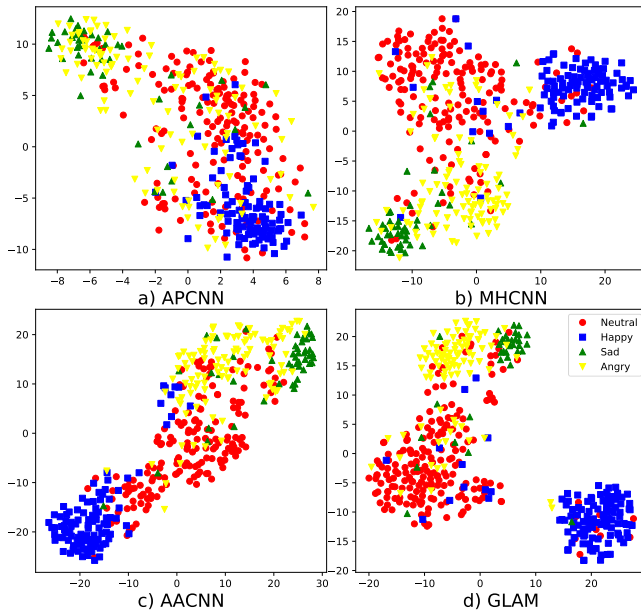

Fig. 2. Confusion matrix on Improvisation dataset. a) and b) represent results of AACNN and GLAM model respectively.

4.2. Results and Discussion

Experiments are conducted on three types of datasets with four common metrics, shown in Table 1. Compared to recent the state-of-the-art approaches, the proposed architecture significantly improves performances on all listed metrics, which demonstrates the effectiveness of global-aware fusion and multi-scale technique.

Meanwhile, GLAM has smaller values of standard deviation for four common metrics which indicates a robust and

stable performance. In detail, we compute a confusion matrix to exploit accuracies for each emotional category, see in Fig. 2. GLAM eliminates confusion sets of *neutral-angry* and *sad-angry* and superiorly increase 6.95% on prediction of *sad* label and 4.52% on prediction of *angry* label. With t-distributed stochastic neighbor embeddings (t-SNE) method [19], we show 2D projections of high-level features in Fig. 3 by using the last second layer output for APCNN, MHCNN and GLAM and applying an additional dense layer to resize large sized output for AACNN. The visualization of APCNN in Fig. 3 a) shows a drastic overlap among four categories. For a MHCNN model shown in Fig. 3 b), the *happy* category is unobviously departed from other categories. For a AACNN model shown in Fig. 3 c), the areas of four categories are successively joint together. In our model, see in Fig. 3 d), the *happy* category is apparently separated from other categories. We see a few *angry* points located in area of *neutral*. There is a better distinction between *sad* and *angry* categories compared to AACNN model.

Table 2. Performance of multi-scale module and global-aware module on Improvisation dataset.

Model	WA	UA
GLAM	**81.18±1.47**	**79.25±1.88**
Multi-scale	80.89±1.43	78.85±1.65
Multi-scale + MHA	80.43±1.53	78.92±1.92
Multi-scale + AA	80.61±1.75	79.09±1.75
MHCNN	76.09±2.01	73.87±2.31
AACNN	78.47±2.42	76.68±3.29

Table 3. Result of α on Improvisation dataset.

α	WA	UA
0	80.44±1.55	78.86±1.65
0.3	81.15±1.73	79.26±1.98
0.5	81.18±1.47	79.25±1.88
0.8	80.97±1.53	78.97±1.85
1	80.96±1.72	79.19±1.92
2	80.80±1.39	79.18±1.66
3	80.61±1.45	78.95±1.58

4.3. Ablation Study

We conduct ablation experiments on the model by removing and replacing modules used in our proposed module. Results are shown in Table 2. By removing the global-aware fusion module, the multi-scale module universally outperforms over all existing methods. By replacing such module with Multi-Head Attention (MHA) [8] or Area Attention (AA) [9], the performance even degrades compared to pure multi-scale module, which indicates effective attention on high dimensional representation globally rather than local fusion.

Selection of α. Without the *mixup* method (i.e. $\alpha = 0$), our model surpasses previous works. Further improvement is realized by using *mixup* method with the optimal result for $\alpha = 0.5$ in Table 3, which indicates the effectiveness of enhancing generalization capability.

Fig. 3. Visualizations of hidden states by commonly-used t-SNE method[19] on improvisation dataset for four models: a) APCNN; b) MHCNN; c) AACNN; d) GLAM.

From above results, we see effectiveness of the multi-scale technique with a global-aware perceiving. Previous works only scale features with fixed length. And current attention-based SER methods are lack of perceiving global information. While, GLAM exploits features at different scales for multiple spanned patterns and attend global information across different features.

5. CONCLUSION

In this paper, we propose a novel architecture, named as GLAM network, which combines a multi-scale module and a global-aware fusion module. The multi-scale module captures different scaled features for various time-spanned patterns and the global-aware module enhances cross-scale communication and attends feature confusion at different scales. In addition, we use mixup method to enhance the generalization of our model. Experiment results demonstrate that our model achieves the state-of-the-art performance on the benchmark dataset IEMOCAP over 4-class emotion classification.

6. REFERENCES

[1] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] Aharon Satt, Shai Rozenberg, and Ron Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *2017 Annual Conference of the International Speech Communication Association*, 2017.

[3] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 International Conference on Platform Technology and Service*, 2017.

[4] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," in *2018 Annual Conference of the International Speech Communication Association*, 2018.

[5] Mustaqeem and Soonil Kwon, "A cnn-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, 2020.

[6] Pengcheng Li, Yan Song, Ian Vince McLoughlin, Wu Guo, and Li-Rong Dai, "An attention pooling based representation learning method for speech emotion recognition," *2018 Annual Conference of the International Speech Communication Association*, 2018.

[7] Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, and Yanhui Tu, "Attention based fully convolutional network for speech emotion recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2018.

[8] Mingke Xu, Fan Zhang, and Samee U. Khan, "Improve accuracy of speech emotion recognition with attention head fusion," in *2020 Annual Computing and Communication Workshop and Conference*, 2020.

[9] Mingke Xu, Fan Zhang, Xiaodong Cui, and Wei Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.

[10] Kun Han, Dong Yu, and Ivan Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *2014 Annual Conference of the International Speech Communication Association*, 2014.

[11] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[12] Xixin Wu, Songxiang Liu, Yuewen Cao, Xu Li, Jianwei Yu, Dongyang Dai, Xi Ma, Shoukang Hu, Zhiyong Wu, Xunying Liu, and Helen Meng, "Speech emotion recognition using capsule networks," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[13] Lorenzo Tarantino, Philip N. Garner, and Alexandros Lazaridis, "Self-attention for speech emotion recognition," in *2019 Annual Conference of the International Speech Communication Association*, 2019.

[14] Hanxiao Liu, Zihang Dai, David R. So, and Quoc V. Le, "Pay attention to mlps," *CoRR*, vol. abs/2105.08050, 2021.

[15] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *2018 International Conference on Learning Representations*, 2018.

[16] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[17] Moataz M. H. El Ayadi, Mohamed S. Kamel, and Fakhri Karray, "Speech emotion recognition using gaussian mixture vector autoregressive models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[18] Weihui Dai, Dongmei Han, Yonghui Dai, and Dongrong Xu, "Emotion recognition and affective computing on vocal social media," *Information & Management*, vol. 52, no. 7, pp. 777–788, 2015.

[19] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.