Improve Accuracy of Speech Emotion Recognition with Attention Head Fusion

Mingke Xu
Computer Science and Technology
Nanjing Tech University
Jiangsu Province, China
201861120004@njtech.edu.cn

Fan Zhang
IBM Watson Group
IBM Massachusette Lab
Littleton, MA
fzhang@us.ibm.com

***Abstract*—Speech Emotion Recognition (SER) refers to the use of machines to recognize the emotions of a speaker from his (or her) speech. In this paper, we propose a multi-head self-attention based attention method to improve the recognition accuracy of SER. We call this method head fusion. We combined this method to implement an ACNN model, using MFCCs extracted from speech and conducted experiments and evaluations on the IEMOCAP dataset. 76.18% of WA and 76.36% of UA were obtained on the improvised part, both better than state-of-the-art.**

***Keywords*-speech emotion recognition; convolutional neural network; self-attention**

I. Introduction

Today, machine speech recognition services such as Auto Speech Recognition (ASR) have been widely used in society. The machine can easily recognize what humans are talking about. Speech Emotion Recognition (SER) also has broad prospects in the field of criminal investigation, medical care, etc. However, for machines, it is still a huge challenge to recognize emotional content in human discourse.

Unlike general ASR, since emotions are subjective and influenced by many factors, how to obtain accurate and recognized high-quality labeled emotional speech data is important in SER. According to R. Altrov *et al.* [1], language and culture have an important influence on the judgment of emotions in speech. Fortunately, the IEMOCAP corpus established by C. Busso *et al.* effectively solved the problem of lack of data in SER [2]. This corpus has been frequently used in the field of SER.

However, the current SER model still does not accurately recognize human emotions. On the IEMOCAP corpus, the state of the art recognition accuracy is 70.17% for weighted accuracy(WA), and 70.85% for unweighted accuracy (UA) [3]. We believe that there is still much room for improvement in this accuracy. In this paper, we propose an approach based on multi-head self-attention and architecture combined with attention mechanism and Convolutional Neural Network (CNN) for SER. It achieved state of the art recognition accuracy on IEMOCAP.

In this paper, our main contributions are as follows:

1) We proposed a method based on multi-head self-attention. We call this method head fusion. Using this method increases the recognition accuracy by about 6% in the IEMOCAP corpus compared to the normal self-attention structure. 2) We implemented an SER model combining CNN and attention and performed experiments on the IEMOCAP corpus, reaching 76.18% of WA and 76.36% of UA, which is state of the art.

II. Related Work

Before the era of deep learning, for SER, researchers mostly use complex hand-crafted features (such as IS09, eGeMaps, etc.) and traditional machine learning methods (such as HMM, SVM, etc.) [4], [5]. In 2014, K. Han *et al.* proposed the first end-to-end deep learning SER model [6]. Later, with the development of deep learning technology, more and more deep learning models for SER were proposed.

Vladimir Chernykh *et al.* proposed a CTC-based RNN network that uses hand-crafted 34-dimensional features for emotional classification [7]. Abdul Malik Badshah *et al.* inputted the spectrogram as features into deep CNN for classification [8]. Xixin Wu *et al.* used the spectrogram and replaced the traditional convolution network with CapsNet, which achieved better classification results [9].

At present, the attention mechanism is being used more and more in the field of deep learning, especially after Google has greatly improved the accuracy of machine translation [10]. In the field of speech recognition, attention mechanism has been used in many tasks such as ASR [11], speaker recognition [12] and SER [3], [13]–[15].

Pengcheng Li *et al.* proposed a mechanism called attention pooling that uses a spectrogram as input to apply top-down and bottom-up attention [13]. Lorenzo Tarantino *et al.* uses a combination of general self-attention and CNN, using a hand-crafted feature set [3]. Ziping Zhao *et al.* combines the Connectionist Temporal Classification(CTC) method and attention mechanism to convert classification problems into transfer problems by adding silent labels [14]. Mingyi Chen *et al.* uses the mel-spectrum, its delta and its second-order delta as inputs, combines CNN and RNN, and uses an attention layer at the end [15].

In addition, some studies combine speech and text for recognition. For example, Seunghyun Yoon *et al.* use ASR to

obtain text, use an attention mechanism to calculate speech-related parts from the text, and use BLSTM for recognition [16].

III. MODEL ARCHITECTURE AND HEAD FUSION

SER is a classification problem that extracting features from human speech and inputting them into traditional machine learning algorithms or deep neural networks for recognition. Emotions are subjective and complex. If the tendency of emotions (positive or negative) is the same, the differences between them are so small that even humans can not easily and accurately distinguish, which brings great difficulties to SER. In order to improve the accuracy of SER, we propose the following methods.

A. Feature Extraction

We use MFCCs as input, an audio feature that is widely used in the field of speech recognition. First, we use a Hanning window with a length of 2048 and a hop length of 512 to perform a short term Fourier transform (STFT) on the audio signal and obtain the power spectrum of the audio signal. Then we use mel filters to map the spectrum to Mel-scale and log to obtain the log Mel-spectrum. Finally, we use a DCT transformation to obtain 13 MFCCs.

B. Model Architecture

Figure 1 shows the overall structure of our model, which consists mainly of a convolutional layer and an attention layer. We use the extracted MFCCs as input and treat this input as an image, using two convolutional layers with a kernel size of (10,2) and (2,8) to extract the horizontal (cross-time) and vertical (cross-MFCC) textures, respectively. After padding, the 8-channel output of each convolutional layer is concatenated to a 16-channel representation. Then 4 convolution layers are applied to generate an 80-channel representation and sent to the self-attention layer. Table I shows the specific settings for the convolutional layers. After each convolutional layer, a Batch Normalization (BN) layer and an activation function Relu are used, and conv2 and conv3 are followed by max-pooling with a kernel size of 2 to reduce the data size.

Table I
THE SPECIFIC SETTINGS FOR THE CONVOLUTIONAL LAYERS.

Name	Settings
Conv1a	kernel size=(10,2),stride=1,in channels=1,out channels=8
Conv1b	kernel size=(2,8),stride=1,in channels=1,out channels=8
Conv2	kernel size=(3,3),stride=1,in channels=16,out channels=32
Conv3	kernel size=(3,3),stride=1,in channels=32,out channels=48
Conv4	kernel size=(3,3),stride=1,in channels=48,out channels=64
Conv5	kernel size=(3,3),stride=1,in channels=64,out channels=80

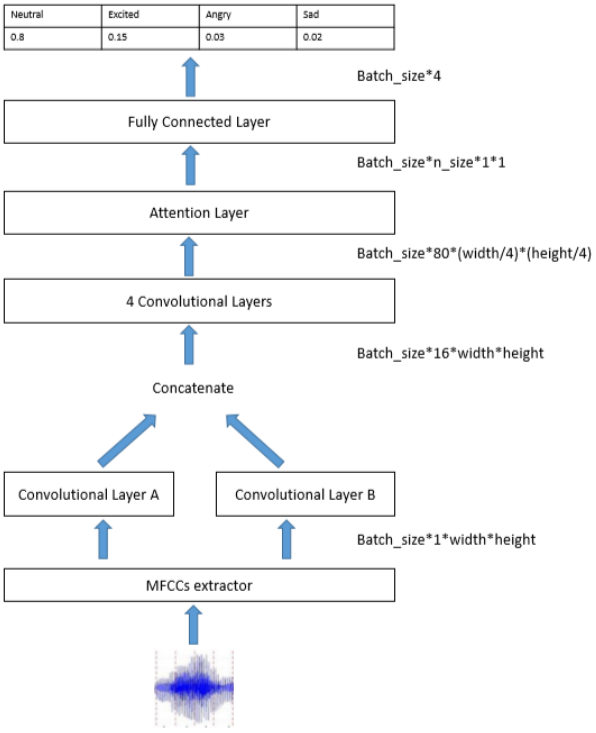


Figure 1. Overall structure of our ACNN model.

C. Self-Attention Layer and Head Fusion

For the 80-channel representation X_{cnn} generated by CNN, we calculate

$$K = W_k * X_{cnn}, Q = W_q * X_{cnn}, V = W_v * X_{cnn}, \quad (1)$$

where W_k, W_q, W_v are trainable parameters. After this we calculate

$$X_{attn} = Softmax(KQ^T)V \quad (2)$$

to obtain X_{attn}——an attention map of X_{cnn}. We use ${X_{attn}}^i$ to represent i_{th} X_{attn} and calculate ${X_{attn}}^i$ by using different parameter sets ${W_k}^i$, ${W_q}^i$, ${W_v}^i$,where $i \in (0, n\_head]$, and each ${X_{attn}}^i$ is called a head. Different from the general self-attention, we superimpose heads to obtain an attention map with multiple points of attention.

$$X_{mattn} = \frac{\sum_0^{n\_head-1} X_{attn}^i}{n\_head} \quad (3)$$

Then we use global average pooling (GAP) to generate a feature point X_{fusion} for this map. In the past computer vision research, GAP has been proved to be an effective method [17]. We call this superposition method head fusion. We set hyperparameter n_head to represent how many heads being fused in a feature point and n_size to represent how many feature points to generate. Finally, we concatenate these points and feed them to the fully connected layer

to obtain the final classification result. Figure 2 shows the working process of head fusion.

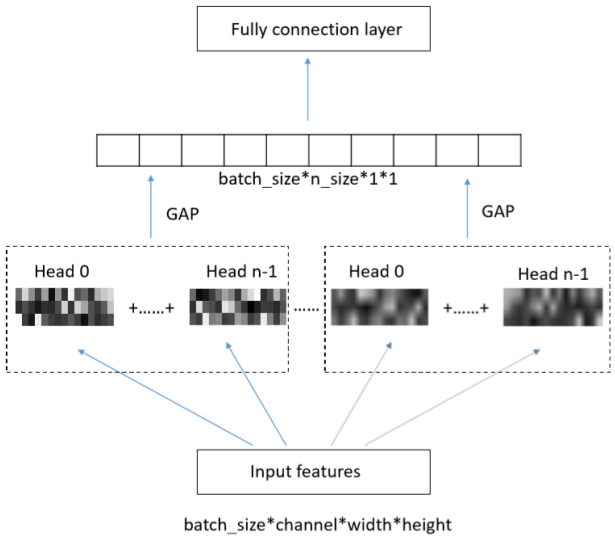


Figure 2. The working process of head fusion.

IV. EXPERIMENTAL EVALUATIONS

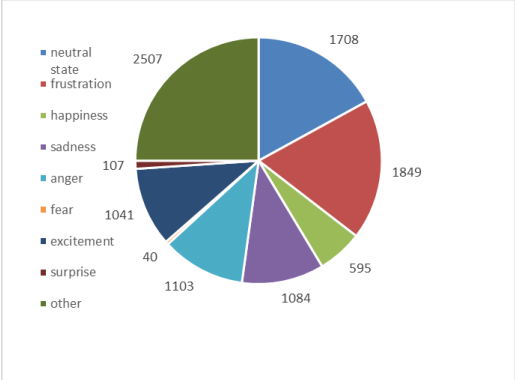
A. Data Set

We use the IEMOCAP corpus as the experimental data set. It is widely used in SER and contains a total of about 12 hours of labeled emotional speech data. It consists of nine emotions: anger, happiness, excitement, sadness, frustration, fear, surprise, other and neutral state. Each utterance is evaluated by an evaluator of 3 or more people, and the utterance will be labeled with the corresponding emotion only if more than half of the evaluators agree, else, it will be labeled with 'other'.

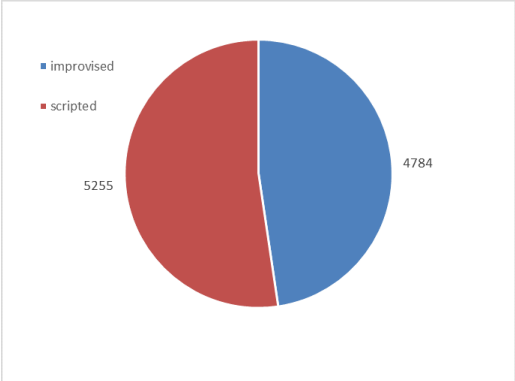
The IEMOCAP corpus is divided into two parts: scripted part and improvised part, that is, the actors perform according to the script and improvisation. In general, the accuracy of the classification of the improvised part is higher than that of the scripted part, because actors do not need to pay attention to the content of the words and express emotions more naturally [3], [13].

Figure 3 shows the distribution of data in the IEMOCAP corpus.

Since the happy class in IEMOCAP is too rare, researchers sometimes choose to use excitement class instead of happy class [3], [7] or merge samples from happy class and excitement class [14], [18]. We selected the improvised part of four emotions (angry, sad, excited and neutral) for our experiment to compare with previous research. Besides, we tested our best model on the scripted part and full corpus to compare with the self-attention architecture in [3].



(a) Distribution of utterances with 9 emotions.



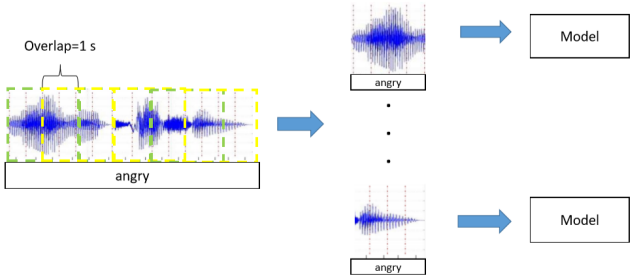
(b) Distribution of utterances improvised or scripted.

Figure 3. Figure (a) shows the distribution of utterances with 9 emotions and Figure (b) shows the distribution of utterances improvised or scripted in the IEMOCAP corpus.

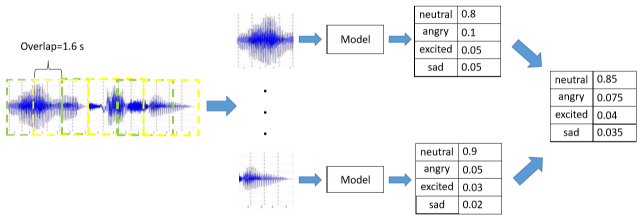
B. Experimental Setup

1) Pre-treatment: For each utterance, we segment it into segments of length 2 seconds and drop the parts that are too short. In the training set, in order to obtain more training data, there is an overlap of 1 second between each segment. These segments are given the label of their source utterance and participate in training as independent data. In the test set, segments from the same utterance are used together, and the results are averaged to obtain predictions. We set the overlap to 1.6 seconds to get better predictions. Figure 4 shows the 2 different ways to pre-treat data in the train set and the test set.

2) Verification Method: To compare with previous studies, we used 5-fold cross-validation, randomly selected 80% data for training and 20% data for testing. We implemented the model with PyTorch, used the cross-entropy loss function, and optimized it with the Adam optimizer. The batch size was set to 32. The initial learning rate was 0.001, weight decay was 1e-6, and the learning rate was manually set to 1/10 for every 10 epochs. We trained the model with 50 epochs on a GTX 1060 GPU and evaluated its WA and UA, saving the best model. For each parameter setting, we used 5



(a) The way of pre-treatment in train set.



(b) The way of pre-treatment in test set.

Figure 4. As shown in Figure (a), in the train set, we segment an utterance into segments of length 2 seconds with an overlap of length 1 second between each segment and use them as independent data in training. As shown in Figure (b), in the test set, we segment an utterance into segments of length 2 seconds with an overlap of length 1.6 seconds between each segment and average the prediction of each segment of a source utterance to obtain the final prediction.

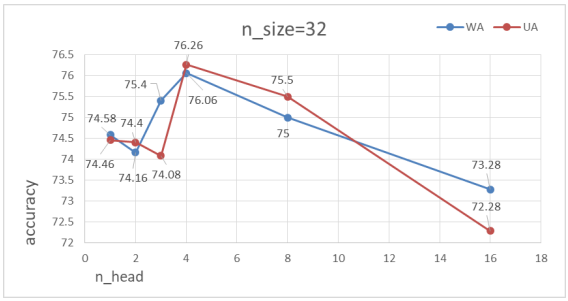
different random seeds for training and testing and averaged the accuracy to reduce the error.

C. Results and Analysis

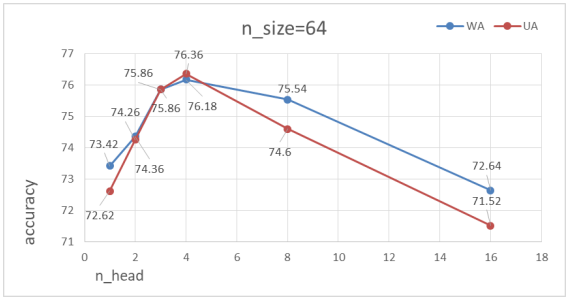
In the experiment, the accuracy is improved compared to the self-attention model without head fusion, and the value of n_head has a significant effect on the experimental results. We find that as n_head increases, the accuracy increases gradually and reaches a maximum at a certain point. Then, if n_head continues to increase, the accuracy will gradually decrease. We believe that the reason for this decline is that the value of n_head is too high, causing some attention points in the map to be placed in too much detail, which will reduce the effect of attention instead.

Figure 5 shows the change in accuracy caused by changing the value of n_head when n_size is set to 32, 64, and 128. When n_size is set to 32 or 64, we obtain the highest accuracy when n_head is set to 4, and when n_size is set to 128, we obtain the highest accuracy when n_head is set to 3.We set n_head to 4 in our final model.

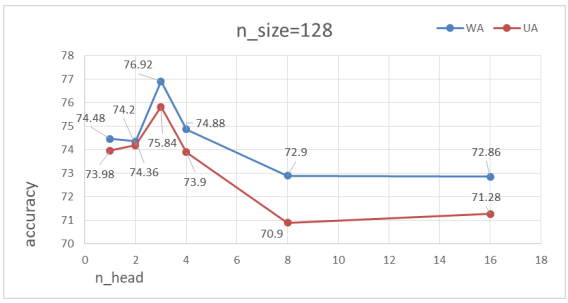
As shown in Figure 6, we also tested the effect of different n_sizes on accuracy with n_head set to 2,4 and 8. When n_head is set to 2, we obtain the highest accuracy when n_size is set to 16, when n_head is set to 4, we obtain the highest accuracy when n_size is set to 64, and when n_head is set to 8, we obtain the highest accuracy when n_size is set to 32. We set n_size to 64 in our final model.



(a) when n_size is set to 32, we obtain the highest accuracy when n_head is set to 4



(b) when n_size is set to 64, we obtain the highest accuracy when n_head is set to 4



(c) when n_size is set to 128, we obtain the highest accuracy when n_head is set to 3

Figure 5. Figure (a) shows the accuracy when n_size is set to 32, Figure (b) shows the accuracy when n_size is set to 64 and Figure (c) shows the accuracy when n_size is set to 128. As shown, when n_size is set to 32 or 64, we obtain the highest accuracy when n_head is set to 4, and when n_size is set to 128, we obtain the highest accuracy when n_head is set to 3.

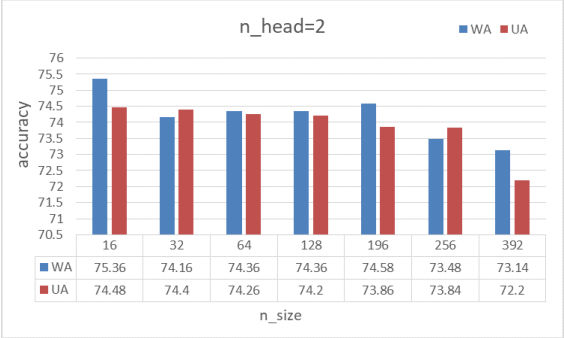
We provide a comparison of the accuracy of previous research and our model in Table II, all of the experiments used the improvised part of IEMOCAP as data set.

To compare with the self-attention architecture in [3], we tested our best model on the scripted part and full corpus of IEMOCAP. The result is shown in Table III.

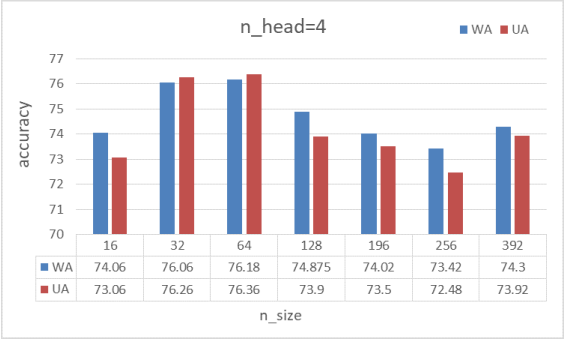
We also performed an ablation study to verify the effect of the attention layer and compare the effects of the number of intermediate convolution layers.Table IV shows the result.

V. Conclusion

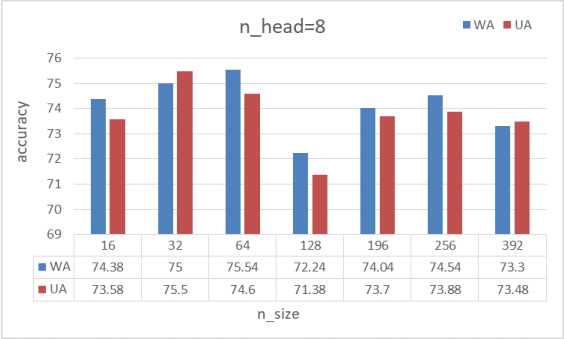
In this paper, we propose an improved mechanism head fusion for SER based on multi-head self-attention and verify



(a) when n_head is set to 2, we obtain the highest accuracy when n_size is set to 16



(b) when n_head is set to 4, we obtain the highest accuracy when n_size is set to 64



(c) when n_head is set to 8, we obtain the highest accuracy when n_size is set to 32

Figure 6. Figure (a) shows the accuracy when n_head is set to 2, Figure (b) shows the accuracy when n_head is set to 4 and Figure (c) shows the accuracy when n_head is set to 8. As shown, when n_head is set to 2, we obtain the highest accuracy when n_size is set to 16, when n_head is set to 4, we obtain the highest accuracy when n_size is set to 64, and when n_head is set to 8, we obtain the highest accuracy when n_size is set to 32.

the role of its parameters. We also implemented an ACNN model, using MFCCs as input features to recognize emotions in speech. Experiments were carried out on the IEMOCAP corpus, using four emotions (angry, sad, excited and neutral) for recognition, which verified the validity of our model.

Table II
COMPARISON OF THE ACCURACY OF PREVIOUS RESEARCH AND OUR MODEL.

Method	WA	UA
Our model	76.18	76.36
Pengcheng Li *et al.* [13]	71.75	68.06
Lorenzo Tarantino *et al.* [3]	70.17	70.85
Ziping Zhao *et al.* [14]	67	69
Gaetan Ramet *et al.* [19]	68.8	63.7
Michael Neumann *et al.* [20]	62.11	/

Table III
COMPARISON OF THE ACCURACY OF MODEL IN [3] AND OUR MODEL.

Method	WA	UA
Our model(improvised)	76.18	76.36
Our model(scripted)	65.9	63.92
Our model(full)	67.28	67.94
Lorenzo Tarantino *et al.* [3](improvised)	70.17	70.85
Lorenzo Tarantino *et al.* [3](scripted)	64.59	50.12
Lorenzo Tarantino *et al.* [3](full)	68.1	63.8

REFERENCES

[1] R. Altrov and H. Pajupuu, "The influence of language and culture on the understanding of vocal emotions," *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, vol. 6, no. 3, pp. 11–48, 2015.

[2] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[3] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," *Proc. Interspeech 2019*, pp. 2578–2582, 2019.

[4] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 2. IEEE, 2003, pp. II–1.

[5] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2010.

[6] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth annual conference of the international speech communication association*, 2014.

[7] V. Chernykh and P. Prikhodko, "Emotion recognition from speech with recurrent neural networks," *arXiv preprint arXiv:1701.08071*, 2017.

[8] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep

Table IV
ABLATION STUDY FOR VERIFYING THE EFFECT OF THE ATTENTION LAYER AND COMPARING THE EFFECTS OF THE NUMBER OF INTERMEDIATE CONVOLUTION LAYERS

Method	WA	UA
1 convolutional layer	63.94	60.14
2 convolutional layers	73.94	74.36
3 convolutional layers	74.68	74.66
4 convolutional layers	76.18	76.36
4 convolutional layers without attention layer	68.7	66.9

convolutional neural network," in *2017 international conference on platform technology and service (PlatCon)*. IEEE, 2017, pp. 1–5.

[9] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu *et al.*, "Speech emotion recognition using capsule networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6695–6699.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[11] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Muller, and A. Waibel, "Very deep self-attention networks for end-to-end speech recognition," *arXiv preprint arXiv:1904.13377*, 2019.

[12] M. India, P. Safari, and J. Hernando, "Self multi-head attention for speaker recognition," *arXiv preprint arXiv:1906.09890*, 2019.

[13] P. Li, Y. Song, I. V. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition." in *Interspeech*, 2018, pp. 3087–3091.

[14] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Attention-enhanced connectionist temporal classification for discrete speech emotion recognition," *Proc. Interspeech 2019*, pp. 206–210, 2019.

[15] M. Chen, X. He, J. Yang, and H. 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.

[16] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2822–2826.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7390–7394.

[19] G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 126–131.

[20] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv preprint arXiv:1706.00612*, 2017.