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

Samee U. Khan
Electrical and Computer Eng
North Dakota State Univ
Fargo, ND
samee.khan@ndsu.edu

Abstract**—Speech Emotion Recognition (SER) refers to the use of machines to recognize the emotions of a speaker from his (or her) speech. SER has broad application prospects in the fields of criminal investigation and medical care. However, the complexity of emotion makes it hard to be recognized and the current SER model still does not accurately recognize human emotions. 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. With this method, an attention layer can generate some attention map with multiple attention points instead of common attention maps with a single attention point. We implemented an attention-based convolutional neural networks (ACNN) model with this method and conducted experiments and evaluations on the Interactive Emotional Dyadic Motion Capture(IEMOCAP) corpus, obtained on improvised data 76.18% of weighted accuracy (WA) and 76.36% of unweighted accuracy (UA), which is increased by about 6% compared to the previous state-of-the-art SER model.**

Index Terms—speech emotion recognition, convolutional neural network, attention mechanism, pattern recognition, machine Learning

I. Introduction

Today, machine speech recognition services such as Automatic Speech Recognition (ASR) have been widely used in society. The machine can easily recognize what humans are talking about. As shown in Fig. 1, similar to ASR, Speech Emotion Recognition (SER) uses machines to recognize humans' emotions when they are talking.

There are two ways to define emotions in SER tasks. The most widely used is to discretely label them as 'happy', 'angry', 'sad' and so on. Although it is intuitive to define emotions discretely, there are about 300 emotional states according to J. O'Connor and G. Arnold's research[21]. It is hard to classify so many emotions. Therefore, the 'palette theory' is agreed with by many researchers, which states that any emotion can be decomposed into primary emotions like the color combined with basic colors. Thus, we only need to classify primary emotions[22]. Another important way is to use a three-dimensional continuous space with parameters like arousal, valence, and potency[23].

Like most of the pattern recognition tasks on computers, an SER task extracts features from raw data and uses these features for classification. In an SER system, the features can be prosodic features like pitch, energy, and intensity. They can also be spectral features like Linear prediction coefficient(LPCs), Perceptual linear prediction coefficients(PLPCs) and Mel-frequency spectrum coefficients(MFCCs)[24]. Sometimes, some kinds of features are combined for better recognition. In this paper, we use MFCCs as the features for recognition.

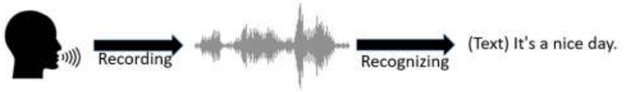
SER has broad prospects in the field of criminal investigation, medical care, etc. An SER system with high accuracy helps judges accurately determine the suspect's psychological condition and make correct decisions, helps psychologists to better understand the patient's mental state and choose the suitable medical treatment and so on. In that, we are committed to improving the accuracy of SER.

Although SER has been considered an important part of Human-Computer Interaction(HCI)[23], there are still many problems that need to be solved, e.g., the lack of data, the difficulty for a machine to describe emotions and etc.

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

However, also due to the subjectivity and uncertainty of emotions, for machines, it is still a huge challenge to recognize emotional content in human speech, *i.e.*, 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]. The WA means the accuracy of all test utterances and UA means the average accuracy of all test classes, which is the evaluation standard commonly used in SER research.

In this paper, our main contributions are as follows:



(a)ASR recognizes what humans are talking about.

The evaluation code is released at github.com/lessonxmk/head_fusion.

978-1-7281-3783-4/20/$31.00 ©2020 IEEE

(b)SER recognizes humans' emotions when they are talking.

Fig. 1. The similarity between ASR and SER

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 general 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 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], also in our work.

Pengcheng Li *et al.* proposed a mechanism called attention pooling that uses a spectrogram as input to apply top-down and bottom-up attention instead of self-attention [13]. Lorenzo Tarantino *et al.* uses a combination of general self-attention and CNN, using a hand-crafted feature set without modification on attention mechanism [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, which is an effective method but the accuracy is still not enough [14]. Mingyi Chen *et al.* uses the Mel spectrogram, its delta and its second-order delta as inputs, combines CNN and Long Short-Term Memory (LSTM), 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]. However, a text combined method depends on the accuracy of ASR.

Different from all the above work, our model only uses audio for recognition, using CNN combined with self-attention as the main structure, and the attention part is modified for higher accuracy than the previous work.

III. Model Architecture and Head Fusion

SER is a classification problem that extracts features from human speech and inputs them into traditional machine learning algorithms or deep neural networks for recognition. Emotion is subjective and complex. Emotions belong to the same category are difficult to be distinguished even by humans. In order to improve the accuracy of SER, we propose the following methods.

A. Feature Extraction

We use Mel-scale Frequency Cepstral Coefficients (MFCCs) as input, an audio feature that is widely used in the field of speech recognition.

Fig. 2 shows the extraction process of MFCCs. 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 spectrogram of the audio signal. Then we use mel filters to map the spectrogram to Mel-scale and take the logs to obtain the log Mel- spectrogram. Finally, we use a discrete cosine transform (DCT) transformation to obtain MFCCs.

B. Model Architecture

Fig. 3 shows the overall structure of our model, which consists mainly of five convolutional layers and an attention layer. We use the extracted MFCCs as input and treat this input as an image. We use two parallel convolutional layers as the first layer with a kernel size of (10,2) and (2,8) to extract the horizontal (cross-time) and vertical (cross-MFCC) textures.

After padding, the 8-channel output of each convolutional layer is concatenated to a 16-channel representation. Then four 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.

C. Self-Attention Layer and Head Fusion

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

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

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

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

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}.

$$X_{mattn} = \frac{\sum_{0}^{n\_head-1} X_{attn}^{i}}{n\_head} \tag{3}$$

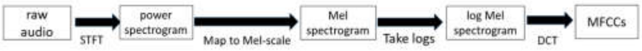


Fig. 2. The extraction process of MFCCs.

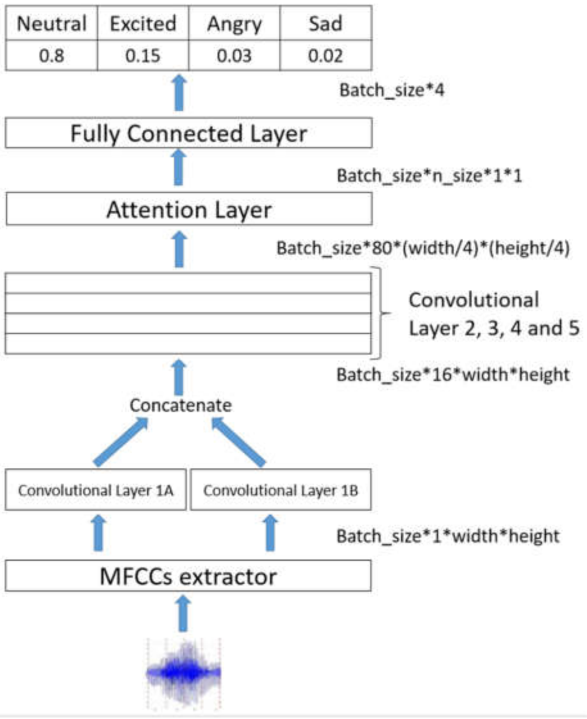


Fig. 3. The overall structure of our ACNN model.

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

Fig. 4 shows the following working process of head fusion. Then we use a global average pooling (GAP) to generate a feature point Xfusion 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.

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 as the corresponding emotion only if more than half of the evaluators agree, else, it will be labeled as ’other’. Therefore, the distribution of emotional states in the IEMOCAP corpus is unbalanced.

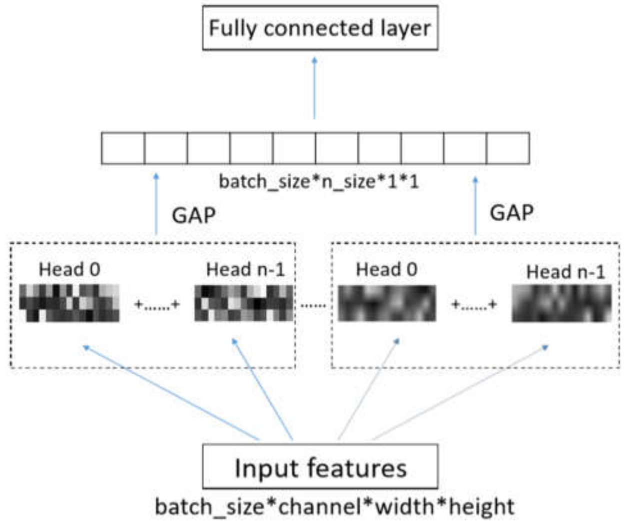


Fig. 4. The working process of head fusion.

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].

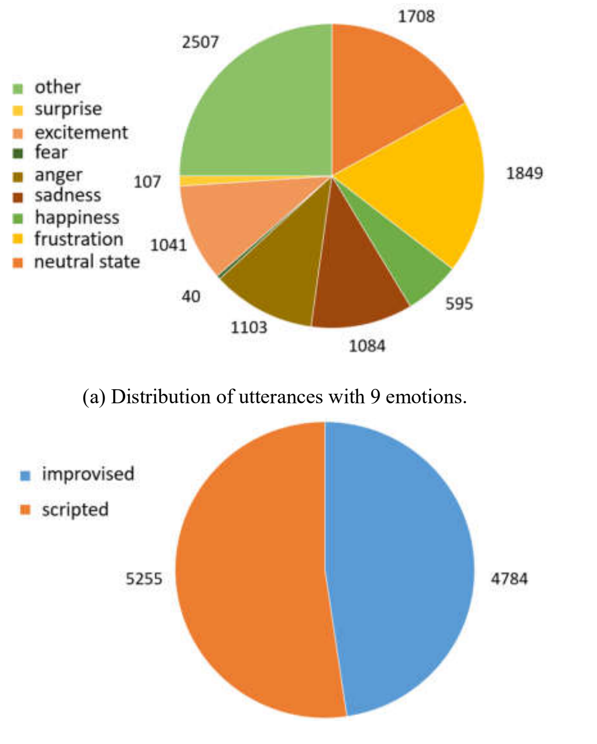
Fig. 5 shows the distribution of data in the IEMOCAP corpus.

Since the happy class in IEMOCAP is too rare, researchers sometimes choose to use an excitement class instead of a 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].

B. Experimental Setup

1) Data Pre-processing

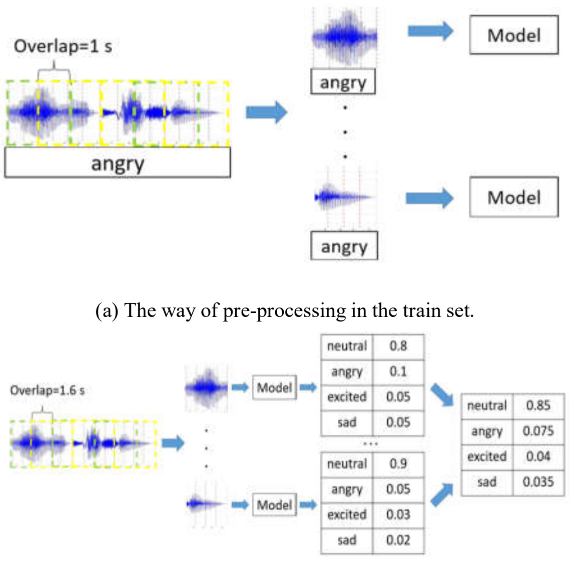
For each utterance, we segment it into segments of length 2 seconds and drop the parts that are too short. In the training set, in order to obtain more training data, there is an overlap of one second between each segment. These segments are given the label of their source utterance and participate in



(a) Distribution of utterances with 9 emotions.

(b) Distribution of utterances improvised or scripted.

Fig. 5. (a) shows the distribution of utterances with nine emotions and (b) shows the distribution of utterances improvised or scripted in the IEMOCAP corpus.



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

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

Fig. 6. As shown in (a), in the train set, we segment an utterance into segments of length 2 seconds with an overlap of length one second between each segment and use them as independent data in training. As shown in (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.

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. Fig. 6 shows the 2 different ways of pre-processing 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 different random seeds for training and testing and averaged the accuracy to reduce the error.

C. *Results and Analysis*

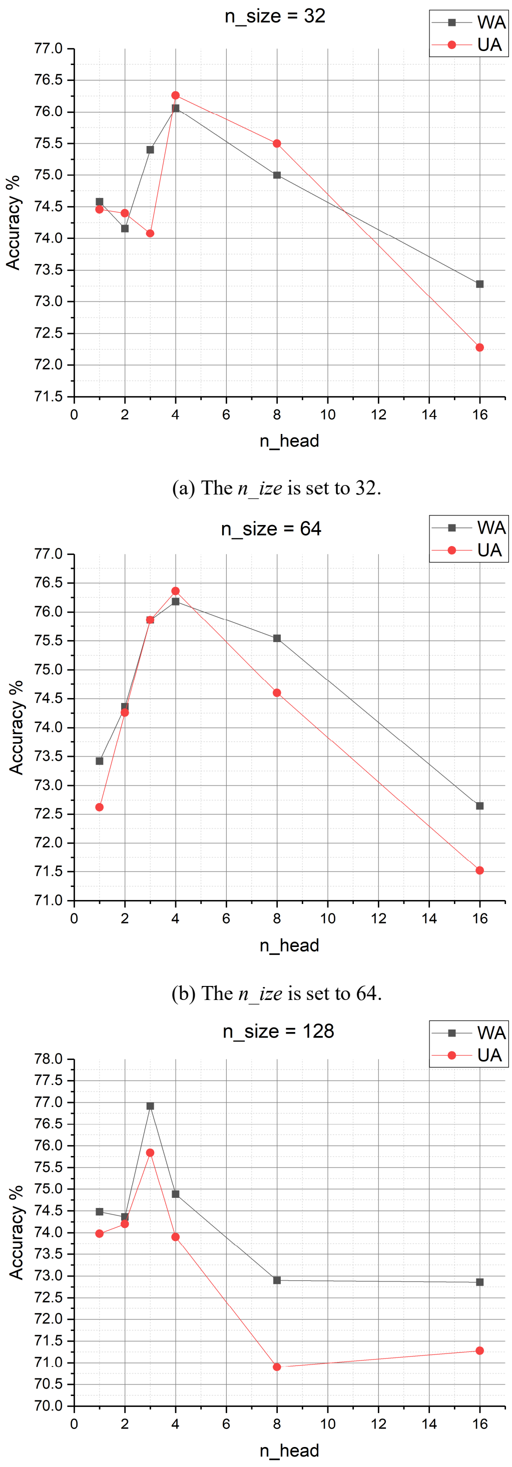
In the experiment, the accuracy is improved compared to the self-attention model without head fusion, and the value of *n_head* has a significant effect on the experimental results. We find that as *n_head* increases, the accuracy increases gradually and reaches a maximum at a certain point. Then, if *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.

Fig. 7 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 Fig. 8, we also tested the effect of different *n_size*s 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.

According to the experimental results, we believe that n_head has a significant impact on accuracy, and n_size has an impact on accuracy but not significant enough. When we fix the value of n_size and change n_head slightly, we can easily obtain the maximum accuracy of a model (in our model, this value is around 76%). However, for n_size, fixing the value of n_head and changing n_size has an effect on the accuracy of the model. In that, we recommend that choose a suitable n_size (such as 32, 64, etc.) and modify n_head to obtain the maximum accuracy when using the head fusion method, instead of focusing on adjusting n_size.

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.

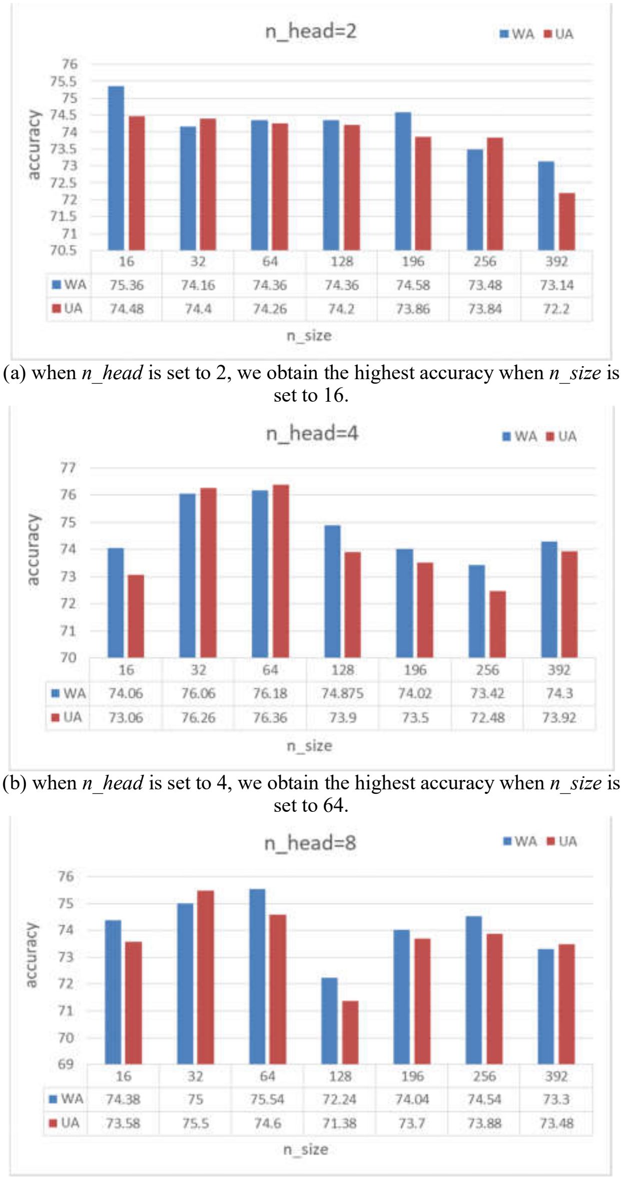


(a) The *n_ize* is set to 32.

(b) The *n_ize* is set to 64.

(c) The *n_ize* is set to 128.

Fig. 7. (a) shows the accuracy when *n_size* is set to 32, (b) shows the accuracy when *n_size* is set to 64 and (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.



(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.

Fig. 8. (a) shows the accuracy when *n_head* is set to 2, (b) shows the accuracy when *n_head* is set to 4 and (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.

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
Tarantino *et al.* [3](scripted)	64.59	50.12
Lorenzo Tarantino *et al.* [3](full)	68.1	63.8

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

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 the role of its parameters. We also implemented an ACNN model, using MFCCs as input features to recognize emotions in speech. Experiments were conducted on the IEMOCAP corpus, using four emotions (angry, sad, excited and neutral) for recognition, which verified the validity of our model. In the future, we plan to use a pre-trained model to obtain the representation of original speech to replace to hand-craft MFCCs. Additionally, we will try to generate more emotional speech for training with generative adversarial networks.

ACKNOWLEDGMENTS

The work of Samee U. Khan was supported by (while serving at) the National Science Foundation (NSF). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

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 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 multihead 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.

[21] J. O'Connor, G. Arnold, Intonation of Colloquial English, second ed., Longman, London, UK, 1973.

[22] El Ayadi M, Kamel M S, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases[J]. Pattern Recognition, 2011, 44(3): 572-587.

[23] Khalil R A, Jones E, Babar M I, et al. Speech Emotion Recognition Using Deep Learning Techniques: A Review[J]. IEEE Access, 2019, 7: 117327-117345.

[24] Peerzade G N, Deshmukh R R, Waghmare S D. A review: Speech emotion recognition[J]. Int. J. Comput. Sci. Eng, 2018, 6: 400-402.