IEICE TRANS. INF. & SYST., VOL.E106–D, NO.5 MAY 2023

LETTER

Speech Emotion Recognition Using Multihead Attention in Both Time and Feature Dimensions

Yue XIE[†a)], ***Member***, **Ruiyu LIANG**[†], **Zhenlin LIANG**[††], **Xiaoyan ZHAO**[†], ***and*** **Wenhao ZENG**[†], ***Nonmembers***

SUMMARY To enhance the emotion feature and improve the performance of speech emotion recognition, an attention mechanism is employed to recognize the important information in both time and feature dimensions. In the time dimension, multi-heads attention is modified with the last state of the long short-term memory (LSTM)'s output to match the time accumulation characteristic of LSTM. In the feature dimension, scaled dot-product attention is replaced with additive attention that refers to the method of the state update of LSTM to construct multi-heads attention. This means that a nonlinear change replaces the linear mapping in classical multi-heads attention. Experiments on IEMOCAP datasets demonstrate that the attention mechanism could enhance emotional information and improve the performance of speech emotion recognition.

key words: *speech emotion recognition, long short-term memory, multi-heads attention, feature enhancement*

1. Introduction

Speech conveys both verbal information and emotional state and is one of the most important types of human communication [1]. Speech emotion recognition (SER) has received increasing interest and has great practical value in human-computer interactions [2] and beyond.

An important issue in the design of an SER system is the extraction of suitable features that efficiently characterize different emotions. Until recently, there was no consensus on the best emotional feature set [2]. With the development of deep learning, neural networks with different architectures have been investigated for SER. In particular, the attention mechanism, which can automatically learn which part of 'the picture' (i. e., the 2D-input) is more important for the final performance in the computer vision field, plays an important role in feature enhancement [3]. Regarding the emotion recognition task, Mirsamadi [4] proposed local attention combined with LSTM for computing weights for frames representing the time series. Based on this work, an attention mechanism was applied on both time and feature dimensions to enhance the output of LSTM. Due to the time accumulation characteristic of LSTM, the attention parameter vector in [4] was replaced with the last state of the LSTM's output.

Manuscript received October 17, 2022.
Manuscript revised January 19, 2023.
Manuscript publicized February 21, 2023.

[†]The authors are with School of Information and Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, P.R.China.
[††]The author is with School of Information Science and Engineering, Southeast University, Nanjing 210096, P.R.China.
a) E-mail: xieyue0109@njit.edu.cn

DOI: 10.1587/transinf.2022EDL8084

Furthermore, dot-product attention, which was first proposed by [5], was extended to multi-heads attention to construct a Transformer architecture for machine translation tasks. Compared with the classical dot-product attention algorithm performing a single attention function, this system mapped attention vectors to multiple new subspaces with different linear projections, which allows the model to jointly attend to information from different representation subspaces at different positions. In the field of emotion recognition, Nediyanchath [6] applied multi-heads attention and position embedding to multitask learning with gender recognition as an auxiliary task to obtain the gender-specific features that influence the emotion characteristics in speech. Li [7] adopted a similar attention approach to obtain a more generalized representation of emotions in a multitask learning framework. Runnan [8] proposed the dilated residual network with multi-heads self-attention for feature learning in SER, which can alleviate the loss of temporal structure and capture the relative dependencies of elements in progressive feature learning.

Although the abovementioned works have successfully applied multi-heads attention to improve the performance of emotion recognition, dot-product attention was used to construct multi-heads attention with linear projection, which means that nonlinear characteristics were ignored. In this study, additive attention [9] with nonlinear projection is employed to build multi-heads attention on the feature dimension. In the time dimension, previous work [7] took the self-attention mechanism as a basic layer that ignores the accumulated characteristic of information in recurrent neural networks, while our previous study [10] demonstrated that better emotion recognition performance could be achieved by taking the last state of LSTM as a reference. In this paper, single-head attention is extended to multi-heads attention to further improve the performance of SER.

2. Classical LSTM

In this paper, LSTM, which was proposed by Schmidhuber [11] for the challenge of learning long-term dependencies in recurrent networks, is adopted to process the temporal information in speech. LSTM has several gates for controlling the flow of information.

$$f_t = \sigma(W_f \times [C_{t-1}, h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \times [C_{t-1}, h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = \tanh(W_C \times [C_{t-1}, h_{t-1}, x_t] + b_C) \quad (3)$$

$$C_t = f_t \bullet C_{t-1} + i_t \bullet \tilde{C}_t \tag{4}$$

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t \bullet \tanh(C_t) \tag{6}$$

where C_{t-1} and h_{t-1} are the cell state and the output of hidden layer at the previous moment, respectively. x_t is the input at the current moment. C_t is the candidate value for updating cell state. f_t, i_t and o_t are the forgetting gate, input gate and output gate, respectively. W_f, W_i, W_C and W_o are the trained weights of forgetting gate, input gate, candidate cell and output gate, and b_f, b_i, b_C and b_o are their biases respectively. $\bullet$ is the Hadamard product. h_t is the output at current moment, while the output of all time steps is denoted as $o_{all\_time} \in R^{B\times M\times N}$ (B represents the size of the batch, M is the number of time steps, and N is the number of hidden units). In classical LSTM, the last moment of output (denoted as $o_{last} \in R^{B\times 1\times N}$, where 1 means the last time step) is selected as the input to full connection layers (or another model that requires fixed-length data as the input) because of the memory ability of LSTM, which causes the accumulated information to be the most abundant in the output of the last moment. However, at the end of speech, there is usually a silent segment that has a native impact on the accumulated information at the last moment. Therefore, an attention mechanism is proposed to weigh the output on the time and feature dimensions.

3. Multi-Heads Attention for LSTM

In previous work [10], single-head attention was applied for LSTM on the time and feature dimensions. In this paper, single-head attention is extended to multi-heads attention to prevent the one-sidedness of a single subspace and enhance the robustness of emotional information to improve performance. In the time dimension, the emotional saturation of speech is different in time segments. The attention mechanism is used to compute weights for frames. In the feature dimension, the ability of different features to distinguish affective categories is different. The attention mechanism is applied to enhance the emotion-relevant features.

3.1 Multi-Heads Attention on the Time Dimension

Since the degree of emotional saturation in each frame is not uniform, the contribution of each frame to the final emotional recognition is different. The degree of contribution can be expressed by the weight coefficients of the frames. Theoretically, the last moment of LSTM networks should obtain a large weight. Therefore, it is taken as a reference to ensure that it can obtain a large weight by using an attention mechanism. In multi-heads attention, the output o_{all_time} and the reference o_{last} are linearly projected into h new subspaces, as shown in Fig. 1. (NOTE: h should be divisible by the number of hidden units N).

$$T\_K_i = o_{all\_time} \times T\_W_{i,k} \tag{7}$$

$$T\_Q_i = o_{last} \times T\_W_{i,q} \tag{8}$$

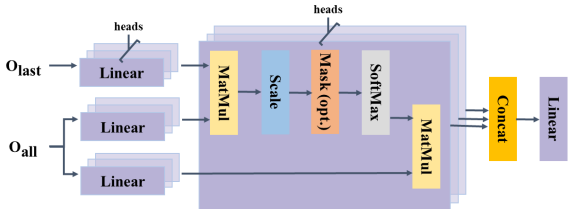

Fig. 1 Multi-heads attention for LSTM in the time dimension

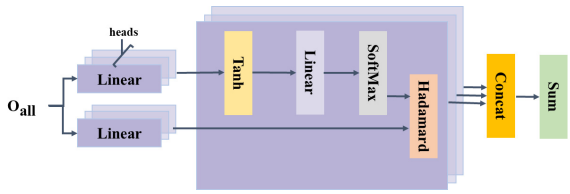

Fig. 2 Multi-heads attention in the feature dimension

where i is the index of subspace that has h in total, $T\_W_{i,k}, T\_W_{i,q} \in R^{N,\frac{N}{h}}$ are weights for linear mappings. Equation (7) completes the mapping of the output o_{all} to the subspace. $T\_Q_i \in R^{B,1,\frac{N}{h}}$ is the new representation of o_{last} in the subspace. The new attention score is calculated by T_K_i and T_Q_i.

$$\mathrm{T}\_score_i = \mathrm{softmax}(\frac{T\_Q_i \times T\_K_i^H}{\sqrt{N}}) \tag{9}$$

$$T\_V_i = o_{all\_time} \times T\_W_{i,v} \tag{10}$$

$$T\_out_i = T\_score_i \times T\_V_i \tag{11}$$

Due to the $T\_out_i \in R^{B,1,\frac{N}{h}}$ is the output of i-th subspace, the score of attention mechanism $T\_score_i \in R^{B,1,M}$ should be multiplied by T_V_i, which is the linear projection of o_{all_time} in the i-th subspace instead of o_{all_time} itself, as shown in Eq. (11). The output of all the subspaces T_out is concatenated and subsequently input into the linear layer with weight matrix T_W_{out}.

$$\mathrm{T}\_Out = Concat(T\_out_1, \ldots, \mathrm{T}\_out_h) \times T\_W_{out} \tag{12}$$

3.2 Multi-Heads Attention on Feature Dimension

It is well known that it is difficult to use single features to accomplish multicategory classification tasks, so multiple features often must be combined to accomplish these tasks. However, the distinguishability of each feature to the target task is not the same. To express the difference among features, an additive attention [9] is calculated on the feature dimension, as shown in Fig. 2

$$F\_K_i = o_{all\_time} \times F\_W_{i,k} \tag{13}$$

$$F\_score_i = \mathrm{softmax}(\tanh(F\_K_i) \times F\_W_{i,score}) \tag{14}$$

$$F\_V_i = o_{all} \times F\_W_{i,v} \tag{15}$$

$$F\_out_i = F\_score_i \bullet F\_V_i \tag{16}$$

where $F\_W_{i,k}, F\_W_{i,v} \in R^{N,\frac{N}{h}}$ are weights of linear projection for the output o_{all_time}. $tanh$ implements nonlinear change in Eq. (14). $F\_W_{i,score} \in R^{\frac{N}{h}\times\frac{N}{h}}$ is the weight to be trained for attention score. The result of the feature dimension F_Out is obtained by concatenating the output of all subspaces.

$$F\_Out = \sum_{time} Concat(F\_out_1, \ldots, F\_out_h) \tag{17}$$

The final output of LSTM is the concatenation of results on the time and feature dimensions. To balance the distribution of data in both dimensions, batch normalization (BN) is performed.

$$Output_{TF} = BN([T\_Out, F\_Out]) \tag{18}$$

4. Experiments and Discussion

The experiments are performed on the IEMOCAP [12] corpus, which consists of five dyadic sessions where actors perform improvisations or scripted scenarios, specifically to represent the emotional expressions. Similar to the reported procedure in state-of-the-art techniques [8], utterances in the "exciting" class are combined with the "happy" class in evaluation to form a four-class database labeled with {happy; angry; sad; neutral}. There are 5,531 utterances in total.

Frame-level features are extracted to retain the temporal information for the attention mechanism on the time dimension [10]. The results are presented as weighted accuracy (WA, overall accuracy on test examples) and unweighted average (UA, average recall over the different emotional categories – a measure commonly used in speech emotion recognition to consider the usually prevalent distributional imbalance across emotion categories.). The proposed network contains 2 LSTM layers with 512 and 256 hidden units, respectively, an attention layer on the time and feature dimensions, a dense layer with 128 hidden units, and a softmax layer. There are approximately 2.6 million parameters.

4.1 Impact of Number of Attention Heads on Performance

To study the influence of the number of attention heads of time dimension (HT) and feature dimension (HF) on performance, the experiments are performed with 0 to 8 attention heads. In Eq. (7), the number of attention heads should be divisible by the number of hidden units, which are set as 256. Classical LSTM, which acts as a baseline in this study could achieve the best UA. Figure 3 shows the experimental results under different combinations of HT and HF. When HT and HF are both zero, classical LSTM is represented without an attention mechanism.

Figure 3 illustrates that the models using the attention mechanism outperform the classical LSTM. As the number of attention heads increases, WA gradually improves. The best WA exceeds 68.0% with 4 and 2 heads on the time and

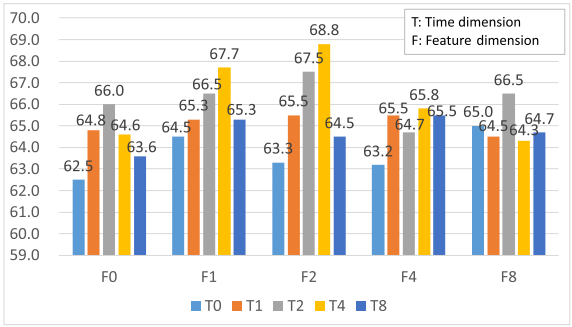

Fig. 3 Impact of the number of heads on performance (WA)

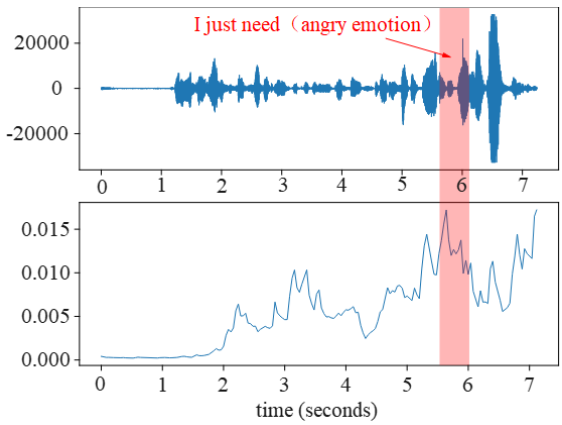

Fig. 4 Attention score on the time dimension

feature dimensions, respectively. However, WA steeply decreases when heads exceed 4 on both the time and feature dimensions. The hidden units may be responsible for this negative phenomenon because of the divisible relationship between the attention heads and the hidden units.

4.2 Attention Score

In this study, only the score on the time dimension is visualized in Fig. 4. The information of the feature dimension is not interpretable after being processed by a neural network, which is one of the disadvantages of deep learning. The speech named 'Ses01F_impro01_M011.wav' (text: 'I don't understand ..., I just need an ID.') is taken as an example from the IEMOCAP corpus. On the top part of Fig. 4, its waveform is plotted. The attention score is shown in the bottom part.

The saturation of emotional expression is different in time, and it is impossible for every frame of speech to have the same emotional components within the duration of speech. Therefore, the purpose of attention mechanism is to find out the difference of emotion in time dimension in this paper. The beginning of speech is usually silent and contains less information. Therefore, the weight coefficient at the beginning of speech should be small (before 1 second, as shown in Fig. 4). As LSTM has the characteristic of temporal accumulation, the weight coefficient should tend

Table 1 Comparison results on IEMOCAP

Models	WA	UA
Baseline	62.5%	62.0%
Mirsamadi [4]	63.5%	58.8%
Nediyanchath [6]	74.1%	64.2%
Li [8]	67.1%	67.4%
Proposed	68.8%	70.0%

to increase gradually over time. At the last moment, the weight should be large. However, not all moments of information are closely related to emotional expression, which leads the weight to decay in the middle of speech (the data at 4 seconds in Fig. 4). In addition, data between 5 and 6 seconds has a relatively large weight, which demonstrates that the expression of emotion is mainly concentrated on certain words (corresponding to the text of 'I just need'). Therefore, the attention mechanism could enhance the information closely related to emotion and suppress distracting information.

4.3 Comparison to State-of-the-Art Approaches

Three related works using an attention mechanism and reporting performance on IEMOCAP are selected for comparison. In Table 1, the baseline is the classical LSTM without an attention mechanism, which obtains a UA of 62.0% UA. Mirsamadi [4] combined local attention with RNN for SER and achieved a UA of 58.8% UA. In [6], the model achieved a UA of 64.2% by using multi-heads attention alone and 70.1% UA by combining auxiliary learning of gender recognition. Li [8] employed multi-heads attention to enhance a dilated residual network and achieved a UA of 67.4% UA for SER. In terms of the application of multi-heads attention, the methods used in [6] and [8] are similar to the application of the time dimension in this study, but the difference is that the last state of LSTM is introduced as a reference instead of pure self-attention. Using this method, 66.0% WA could be achieved (shown in Fig. 4 when HF=0 and HT=2), which is better than [6]. Although [6] had the highest WA, its UA is relatively low, which reflects the imbalance of the sample number. Compared with [8], who reported a model with 9.9 million parameters, the proposed model has fewer parameters (2.6 million) and achieves higher performance in terms of WA (68.8%) and UA (70.0%), when HF=2 and HT=4.

5. Conclusions

This study extends the single attention function to multi-heads attention in both time and feature dimensions for speech emotion recognition. This process could prevent the one-sidedness of a single subspace and enhance the robustness of emotional information to improve performance. Multiple linear subspaces can alleviate the interlacing problem of different emotional information. Experiments demonstrate that the optimal number of multi-heads is highly dependent on the hidden units. The proposed method not only improves the performance of emotion recognition, but also solves the imbalance between UA and WA. In future work, the proposed model could be applied to multitask learning and multimodel to further improve performance of SER.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (No. 62001215, No. 51908285), Introduce Talent Research Start-up Fund of Nanjing Institute of Technology (No. YKJ201977).

References

[1] Y.B. Singh and S. Goel, "Survey on human emotion recognition: Speech database, features and classification," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), pp.298–301, 2018.

[2] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," Artificial Intelligence Review, vol.43, no.2, pp.155–177, Feb. 2015.

[3] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The Application of Two-level Attention Models in Deep Convolutional Neural Network for Fine-grained Image Classification," 2015 IEEE Conference on ComputerR Vison and Pattern Recognition (CVPR), pp.842–850, 2015.

[4] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural nerworks with local attention," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.2227–2231, 2017.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in Neural Information Processing Systems (NIPS 2017), 2017.

[6] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.7179–7183, 2020.

[7] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," INTERSPEECH 2019, Graz, Austria, pp.2803–2807, 2019.

[8] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6675–6679, 2019.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, San Diego, CA, United States, 2015.

[10] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based LSTM," IEEE/ACM Trans. Audio, Speech, Language Process., vol.27, no.11, pp.1675–1685, 2019.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol.9, no.8, pp.1735–1780, 1997.

[12] 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, pp.335–359, 2008.