Journal of Southeast University (English Edition) Vol. 38, No. 2, pp. 103 - 109 June 2022 ISSN 1003—7985

Multi-head attention-based long short-term memory model for speech emotion recognition

Zhao Yan[1] Zhao Li[1] Lu Cheng[1] Li Sunan[1] Tang Chuangao[2] Lian Hailun[1]

([1]School of Information Science and Engineering, Southeast University, Nanjing 210096, China)
([2]School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China)

Abstract: To fully make use of information from different representation subspaces, a multi-head attention-based long short-term memory (LSTM) model is proposed in this study for speech emotion recognition (SER). The proposed model uses frame-level features and takes the temporal information of emotion speech as the input of the LSTM layer. Here, a multi-head time-dimension attention (MHTA) layer was employed to linearly project the output of the LSTM layer into different subspaces for the reduced-dimension context vectors. To provide relative vital information from other dimensions, the output of MHTA, the output of feature-dimension attention, and the last time-step output of LSTM were utilized to form multiple context vectors as the input of the fully connected layer. To improve the performance of multiple vectors, feature-dimension attention was employed for the all-time output of the first LSTM layer. The proposed model was evaluated on the eNTERFACE and GEMEP corpora, respectively. The results indicate that the proposed model outperforms LSTM by 14.6% and 10.5% for eNTERFACE and GEMEP, respectively, proving the effectiveness of the proposed model in SER tasks.

Key words: speech emotion recognition; long short-term memory (LSTM); multi-head attention mechanism; frame-level features; self-attention

DOI: 10.3969/j.issn.1003-7985.2022.02.001

Speech emotion recognition (SER) plays a significant role in many real-life applications, such as human-machine interaction[1] and computer-aided human communication. However, it is challenging to make machines fully interpret emotions embedded in speech signals owing to the subtlety and vagueness in spontaneous emotional expressions[2-3]. Despite the wide use of SER in applications, its performance remains far less competitive in comparison with those of human beings, and the recognition process still suffers from the local optima trap. Therefore, it is essential to further enhance the performance of automatic SER systems.

Deep learning networks have shown great efficiency in dealing with SER tasks, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), which brings a great improvement in the recognition accuracy[4]. The attention mechanism is also utilized in neural networks to dynamically focus on certain parts of the input. Mirsamadi et al.[5] introduced the local attention mechanism to an RNN to focus on the emotionally salient regions of speech signals. Statistical features were used for the study. Tarantino et al.[6] proposed a new windowing system with the self-attention mechanism to improve the SER performance. These studies[5-6] follow the traditional method of using low-level descriptors as the input. Recently, spectrograms have gained considerable attention as the input feature. For instance, Li et al.[7] adopted the self-attention mechanism for the salient periods of speech spectrogram. Although researchers paid considerable attention to deep networks, the input features are mainly extracted from the time dimension.

Many studies focus on exploring multiple dimensions' feature vectors. Xie et al.[8] proposed a weighting algorithm based on time and feature-dimension attention for the long short-term memory (LSTM) output, which could significantly improve the SER performance. Li et al.[9] combined a dilated residual network and multi-head self-attention to model inner dependencies. For the above algorithms[8-9], the last time-step output of models is used as the input of the next layer. The studies indicate that parallel multiple feature vectors help improve the SER performance. Moreover, the attention mechanism has demonstrated great performance for SER tasks[5-9] and has been used in combination with deep neural networks.

In this research, an improved multi-head attention LSTM model is proposed to overcome the above-mentioned barriers and improve the SER performance. Multi-head time-dimension attention (MHTA) has the ability to jointly attend to information from different representation subspaces at different positions[10]. Deep network outputs are performed in parallel through the attention function and then concatenated for the final values. Compared with a single-head attention output, concatenated values contain various context vectors, which are weighted by different attention functions. The transformer based on

Received 2021-11-15, **Revised** 2022-02-05.

Biographies: Zhao Yan (1993—), male, Ph. D. candidate; Zhao Li (corresponding author), male, doctor, professor, zhaoli@seu.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 61571106, 61633013, 61673108, 81871444).

Citation: Zhao Yan, Zhao Li, Lu Cheng, et al. Multi-head attention-based long short-term memory model for speech emotion recognition [J]. Journal of Southeast University (English Edition), 2022, 38(2): 103 - 109. DOI: 10.3969/j.issn.1003-7985.2022.02.001.

the multi-head attention mechanism is introduced to the pre-training model, named bidirectional encoder representations from transformers (BERT)[11], becoming one of the most successful models for natural language processing. The success of pre-training by BERT makes multi-head attention widely being used for various fields of speech, such as speech recognition. Lian et al.[12] employed the multi-head attention layer to predict the unsupervised pre-training for the Mel spectrum. Tian et al.[13] introduced multi-head attention into the RNN-transducer structure and achieved excellent results. However, the above-mentioned models did not use the mechanism for mining the temporal relations from the LSTM output. Previous studies mainly focused on directly utilizing multi-head attention layers for pre-training and improving the SER performance. From this point of view, the values contain more information of the salient speech region. The all-time output and last time-step output are utilized for the MHTA calculation. Moreover, SER is not just decided by the output of the MHTA but also by different representations from other aspects. Therefore, the output of feature-dimension attention and the last time-step output of LSTM are introduced to the final context vector. The information loss always exists during the backpropagation of the traditional deep learning network. The residual neural network[14-15] helps solve this problem by connecting the previous layer output with the subsequent layer output directly. Inspired by the idea, feature-dimension attention for the all-time output of the first LSTM layer is employed to select useful information. Finally, the context vector is fed to the fully connected layer. Experiments conducted on the eNTERFACE and GEMEP corpora demonstrate the effective performance of the proposed model.

1 Proposed Method

1.1 Frame-level feature extraction

The openSMILE[16] features proposed by Schuller et al.[17] are the most widely used speech features for SER. In this research, to keep the uniformity and coherence of the previous work[8], the same frame-level features are used.

1.2 MHTA mechanism

In this structure, the LSTM layer for processing the time series samples with the variable length is used. The LSTM network, which is proposed by Hochreiter and Schmidhuber[18], models time series sequences and generates a high-level representation. It has the ability to extract features automatically and hierarchically. To maintain continuity and consistency of the previous work[8], which demonstrates the effectiveness of the attention gate, the double-layer LSTM with the attention gate was used in the structure. The LSTM's output can be described as a matrix composed of time steps and feature data. Therefore, neural components are required to learn hidden information between the time steps and features of the output representations. In this paper, improved attention is introduced as the core mechanism for computing the representations' time and feature relationship.

Vaswani et al.[10] introduced an attention function as mapping a query and a set of key-value pairs to an output. The output was computed by weighted values, where the weight was calculated by the query with the corresponding key. Instead of applying a single attention function, Vaswani found it useful to apply multi-head attention functions for queries, keys, and values. The neural network utilizes the multi-head attention algorithm to consume hidden information from different subspaces, which can significantly improve the performance compared with single-head attention.

$$\text{Multi-head}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\boldsymbol{h}_1, \boldsymbol{h}_2, \dots, \boldsymbol{h}_n)\boldsymbol{W}_{\text{O}} \quad (1)$$

$$\boldsymbol{h}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_{i,Q}, \boldsymbol{K}\boldsymbol{W}_{i,K}, \boldsymbol{V}\boldsymbol{W}_{i,V}) \quad i \in [1, n] \quad (2)$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are the query, key, and value vectors; $\boldsymbol{W}_i$ is the parameter matrix for mapping into subspaces; $\boldsymbol{W}_{\text{O}}$ is the mapped-back parameter matrix; n is the number of attention heads.

The frame-level features were selected for SER. Because the frame-level features contain the time and feature relationship of the LSTM output, the attention function applied to the time and feature dimensions helps the model to improve performance. A previous work[8] used the attention mechanism on the time and feature dimensions to achieve state-of-the-art performance for emotion recognition. The attention weighting for the time-dimension calculation is

$$\boldsymbol{S}_{\text{t}} = \text{softmax}(\boldsymbol{O}_{\text{m}}(\boldsymbol{O}_{\text{a}}\boldsymbol{W}_{\text{t}})^{\text{T}}) \quad (3)$$

$$\boldsymbol{O}_{\text{t}} = \boldsymbol{S}_{\text{t}}\boldsymbol{O}_{\text{a}} \quad (4)$$

where $\boldsymbol{O}_{\text{m}} \in \mathbf{R}^{B\times 1\times Z}$ is the last time output and $\boldsymbol{O}_{\text{a}} \in \mathbf{R}^{B\times F\times Z}$ is the all-time output; B is the batch size; F is the number of time steps; Z is the feature's dimension; $\boldsymbol{S}_{\text{t}} \in \mathbf{R}^{B\times 1\times F}$ is the attention score of the time dimension; $\boldsymbol{O}_{\text{t}}$ is the output of the time-dimension attention layer, which is subsequently fed into the fully connected layer.

Theoretically, the multi-head attention algorithm, intended to project the LSTM output into different subspaces for hidden information with different dimensions, could achieve a good performance. Moreover, for each head, the decreased dimensions help to keep the calculation amount similar to that of the single-head attention. The last time-step output of LSTM accumulates the greatest amount of information because of the memory ability of the LSTM network. By using the all-time output of LSTM and last time-step output, the keys, values, and queries are computed as

$$\boldsymbol{K}_i = \boldsymbol{W}_{i,\text{K}}\boldsymbol{O}_{\text{a}} + \boldsymbol{b}_{i,\text{K}} \quad (5)$$

$$\boldsymbol{V}_i = \boldsymbol{W}_{i,\mathrm{V}}\boldsymbol{O}_a + \boldsymbol{b}_{i,\mathrm{V}} \tag{6}$$

$$\boldsymbol{Q}_i = \boldsymbol{W}_{i,\mathrm{Q}}\boldsymbol{O}_m + \boldsymbol{b}_{i,\mathrm{Q}} \tag{7}$$

where $\boldsymbol{K}_i \in \mathbf{R}^{B\times F\times \frac{Z}{n}}$, $\boldsymbol{V}_i \in \mathbf{R}^{B\times F\times \frac{Z}{n}}$, and $\boldsymbol{Q}_i \in \mathbf{R}^{B\times 1\times \frac{Z}{n}}$ are the key, value, and query for a subspace, respectively; $\boldsymbol{W}_i \in \mathbf{R}^{Z\times \frac{Z}{n}}$ is the parameter matrices; $\boldsymbol{b}_i \in \mathbf{R}^{\frac{Z}{n}}$ is the bias.

Next, the calculated keys, values, and queries are utilized to compute the corresponding attention scores and attention output. The calculations are as follows:

$$\boldsymbol{s}_i = \mathrm{softmax}(\boldsymbol{Q}_i\boldsymbol{K}_i^{\mathrm{T}}) \tag{8}$$

$$\boldsymbol{o}_{\mathrm{mt}i} = \boldsymbol{s}_i\boldsymbol{V}_i \tag{9}$$

$$\boldsymbol{O}_{\mathrm{mt}} = \mathrm{Concat}(\boldsymbol{o}_{\mathrm{mt}1}, \boldsymbol{o}_{\mathrm{mt}2}, \ldots, \boldsymbol{o}_{\mathrm{mt}n}) \tag{10}$$

where $\boldsymbol{s}_i$ is the multi-head attention score on the time dimension and $\boldsymbol{o}_{\mathrm{mt}i}$ is its output for each subspace; $\boldsymbol{O}_{\mathrm{mt}}$ is the final output of the MHTA layer. After the output from all the subspaces is obtained, outputs from their corresponding subspaces are concentrated, which will be fed into the fully connected layer.

1.3 Multiple-context-vector generation

For SER, the features exhibit different influences. To classify the feature difference, the feature-dimension attention mechanism is applied in this work. The feature-dimension attention used in the model helps to relieve the overfitting problem caused by the time-dimension multi-head attention algorithm. The feature weighting is calculated as follows:

$$\boldsymbol{s}_{\mathrm{f}} = \mathrm{softmax}(\tanh(\boldsymbol{O}_{\mathrm{m}}\boldsymbol{w}_{\mathrm{f}})\boldsymbol{u}_{\mathrm{f}}) \tag{11}$$

$$\boldsymbol{O}_{\mathrm{f}} = \sum \boldsymbol{s}_{\mathrm{f}}\boldsymbol{O}_{\mathrm{a}} \tag{12}$$

where $\boldsymbol{w}_{\mathrm{f}}$ and $\boldsymbol{u}_{\mathrm{f}}$ are the training parameters. The feature-dimension attention score $\boldsymbol{s}_{\mathrm{f}}$, which is different from each other, could indicate the effect of different features. Next, the summation over the time frames is calculated. The output $\boldsymbol{O}_{\mathrm{f}}$ represents the statistical value of the time-dimension features.

Finally, the last time-step output ($\boldsymbol{O}_{\mathrm{ls}}$) is chosen, which accumulates the greatest amount of information as parts of the final output. The final output consists of three different parallel characterizations. After the context vector is calculated, they are put through the unsqueeze function.

$$\boldsymbol{O}_{\mathrm{tfl}} = \mathrm{Concat}(\boldsymbol{O}_{\mathrm{mt}}, \boldsymbol{O}_{\mathrm{f}}, \boldsymbol{O}_{\mathrm{ls}}) \tag{13}$$

$$\boldsymbol{O}_{\mathrm{ap}} = \mathrm{Averagepooling}(\boldsymbol{O}_{\mathrm{tfl}}) \tag{14}$$

However, because a double-layer LSTM structure is used in this study, the LSTM layer may discard vital information during the process. Therefore, the all-time output of the first LSTM layer ($\boldsymbol{O}_{\mathrm{af}}$) is taken into consideration. The modified output calculation is

$$\boldsymbol{O}_{\mathrm{mo}} = \mathrm{Concat}(\boldsymbol{O}_{\mathrm{mt}}, \boldsymbol{O}_{\mathrm{f}}, \boldsymbol{O}_{\mathrm{ls}}, \boldsymbol{O}_{\mathrm{af}}) \tag{15}$$

Another problem arises because $\boldsymbol{O}_{\mathrm{ls}}$ and $\boldsymbol{O}_{\mathrm{af}}$ could contain the same information, which leads to information redundancy. This situation is not expected to happen because it may have a bad influence on the effectiveness and performance of the model. To avoid such a situation, the feature-dimension attention mechanism is applied for the first LSTM layer's all-time output to screen for useful information.

$$\boldsymbol{s}_{\mathrm{a}} = \mathrm{softmax}(\tanh(\boldsymbol{O}_{\mathrm{af}}\boldsymbol{w}_{\mathrm{f}})\boldsymbol{u}_{\mathrm{f}}) \tag{16}$$

$$\boldsymbol{O}_{\mathrm{al}} = \sum \boldsymbol{s}_{\mathrm{a}}\boldsymbol{O}_{\mathrm{af}} \tag{17}$$

Finally, the multiple context vectors are calculated and used as the input of the fully connected layer.

$$\boldsymbol{O}_{\mathrm{c}} = \mathrm{Concat}(\boldsymbol{O}_{\mathrm{mt}}, \boldsymbol{O}_{\mathrm{f}}, \boldsymbol{O}_{\mathrm{ls}}, \boldsymbol{O}_{\mathrm{al}}) \tag{18}$$

The new context vector ($\boldsymbol{O}_{\mathrm{c}}$) not only provides inherent information from time and feature dimensions but also utilizes the last time-step output as auxiliary information for SER. It can strengthen the key information and ignore irrelevant information to generate a highly effective feature representation. Fig. 1 shows the proposed multi-head attention-based LSTM structure.

2 Experiment and Analysis

2.1 Database

The proposed model is evaluated on eNTERFACE[19] and GEMEP[20] corpus. The eNTERFACE dataset contains 42 subjects (34 male and 8 female). The audio sample rate was 48 MHz, in an uncompressed stereo 16-bit format, with an average duration of 3.5 s. In this research, 1 260 valid speech samples are used for the evaluation, where 260 samples are used as the test set.

GEMEP is a French-content corpus with 18 speech emotional categories, including 1 260 utterance samples. Twelve labeled classes are selected: relief, amusement, despair, pleasure, anger, panic, interest, joy, irritation, pride, anxiety, and sadness. Therefore, 1 080 samples are used from the chosen categories, where 200 samples are selected as the test set randomly.

2.2 Experimental setup

In this section, the proposed model is compared with several baselines: 1) LSTM; 2) LSTM with time-dimension attention; 3) LSTM with MHTA. For experiments performed on the same database, the parameters are kept the same for the LSTM layer.

The input dimension is $[128, t, 93]$, where 128 is the number of batch sizes, t is the frame number, and 93 is the number of extracted features. The output dimension is $[128, c]$, where c represents the number of emotion categories in databases. For the double-layer LSTM, the first layer has 512 hidden units, while the second layer has 256 hidden units. To ensure the dependability and reliability of the experiments, other parameters are kept the same.

As SER is a classification task, the unweighted average recall (UAR) is used as the evaluation metric.

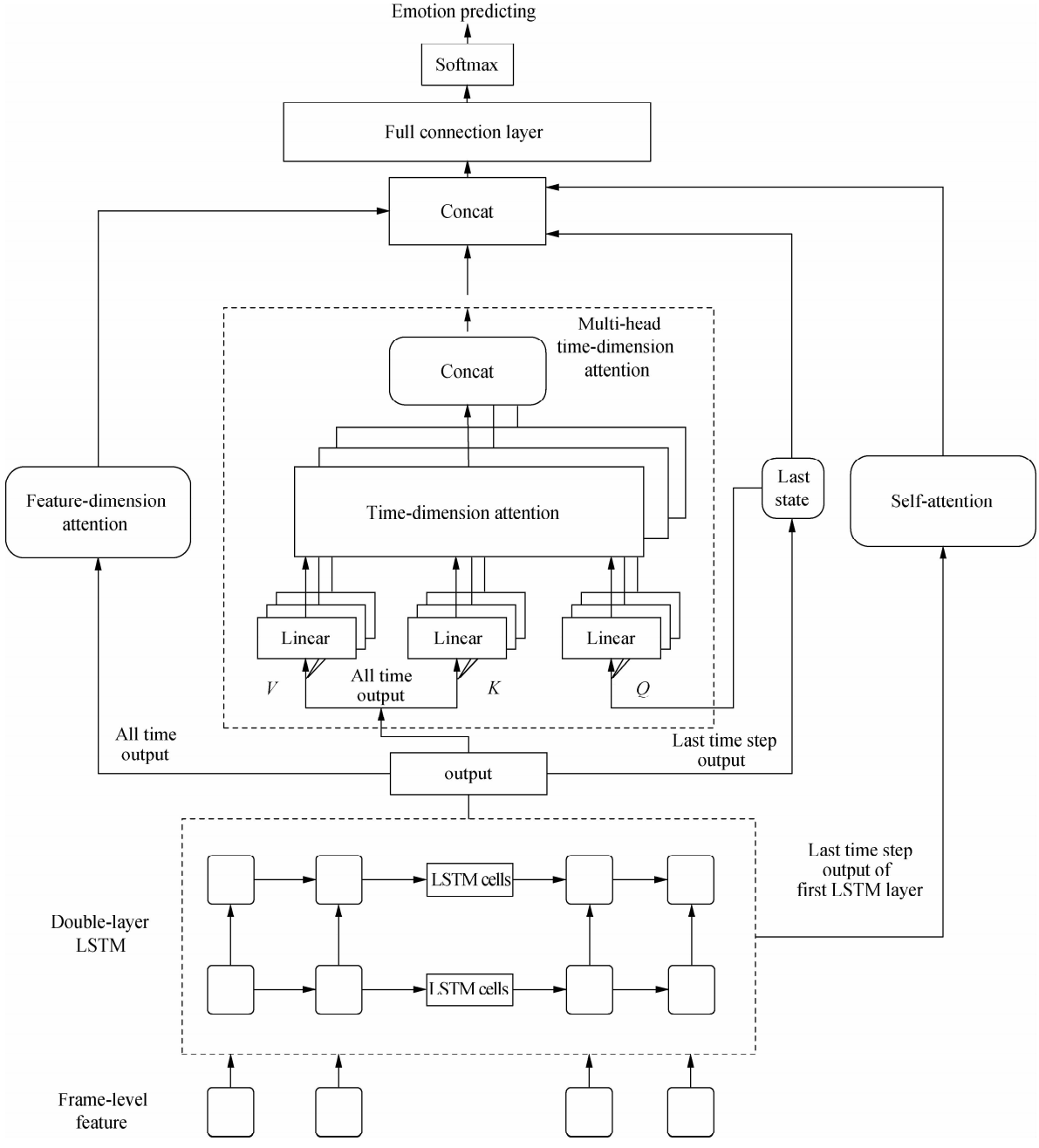

Fig. 1 Proposed improved multi-head attention structure

2.3 Results and discussion

Experiments are conducted to verify the effectiveness of the proposed multi-head attention mechanism. Tab. 1 presents the results of the LSTM model and time-dimension attention-based LSTM models. Compared with the LSTM model, the time-dimension attention LSTM obtains a recognition accuracy of 83.8% and has an 8.0% improvement on the eNTERFACE corpus. For the GEMEP corpus, LSTM with the attention mechanism also obtains an increase of 4.5%. Tab. 2 exhibits the results of the models with three context vectors. The results prove that the LT8 model outperforms other models. The results show a tendency of increasing first and then decreasing, which indicates that the model achieves its boundary when it has eight heads.

Tab. 1 Results of the time-dimension attention LSTM models %

Model	Backbone	M	Accuracy	
			eNTERFACE	GEMEP
L0	LSTM	0	75.8	48.5
L1	LSTM	1	83.8	53.0
L2	LSTM	2	87.3	49.0
L4	LSTM	4	87.7	48.5
L8	LSTM	8	88.1	45.0
L16	LSTM	16	85.8	

Note: M is the head number of MHTA.

Tab. 2 Results of the models with three context vectors %

Model	Backbone	M	Accuracy	
			eNTERFACE	GEMEP
LT1	LSTM	1	86.9	54.0
LT2	LSTM	2	88.1	57.0
LT4	LSTM	4	88.5	57.5
LT8	LSTM	8	89.6	58.0
LT16	LSTM	16	89.2	56.0

Furthermore, the recognition accuracy increases along with the increase in the time-dimension attention head number (less than eight) and then decreases. The reason is that projecting into subspaces has its boundary. Good results cannot always increase by simply increasing the head numbers. This tendency indicates that the multi-head attention is effective for the eNTERFACE corpus. The accuracy of the proposed model is better than that of LSTM, and the curves seem to increase together with the increase in the head number. When the head number is equal to eight, the proposed model achieves the best recognition accuracy of 89.6% and 58.0% in the eNTERFACE and GEMEP corpora, respectively.

Multitask learning[7] has proven its effectiveness for speech recognition. Compared with multitask learning, multiple context vectors are used to determine SER results. Although time-dimension attention would complicate the model, the classification is dependent not only on the context vector of the multi-head attention layer but also on other context vectors. Finally, the output of the feature-dimension attention layer and last time-step output together are composed with the context vector of MHTA to form the final context vector.

In this paper, multiple context vectors are proposed to analyze speech emotions. To evaluate the effectiveness of the proposed method, several experiments are conducted. As several studies have proven the effectiveness of the time-feature attention mechanism[8] and skip connection structure[21] for SER tasks, the performance of the proposed model is compared with the attention LSTM network. As the results of multi-head LSTM models hint that the models achieve the best performance when the head number is equal to eight, it is applied for the attention-based models with different context vectors. Model performance comparisons against other techniques are presented in Tab. 3. The experimental results indicate that the proposed model outperforms LSTM by 14.6% and 10.5% for eNTERFACE and GEMEP, respectively. The UARs decrease in the eNTERFACE and GEMEP databases when the all-time output of the first LSTM layer is employed as an additional context vector for the fully connected layer. This situation may be a result of the all-time output of the first LSTM layer that provides too much redundant information for the model. It makes the fully connected layer input less effective. To solve this problem, the assumption is to employ the feature-dimension attention mechanism for the all-time output of the first LSTM layer to select useful inherent information.

Tab. 3 Results of the LSTM models with various context vectors %

Model	Backbone	M	Accuracy		
			Output	eNTERFACE	GEMEP
L0	LSTM	0		75.8	48.5
L8	LSTM	8		88.1	45.0
LT8	LSTM	8	$\boldsymbol{O}_{\text{tfl}}$	89.6	58.0
LM8	LSTM	8	$\boldsymbol{O}_{\text{mo}}$	88.1	53.0
LC8	LSTM	8	$\boldsymbol{O}_{\text{c}}$	90.4	59.0

The proposed model is compared with other methods. The local attention mechanism[5] is re-implemented for SER. As CNN networks have been widely used for SER tasks, CNN networks[7, 22] are re-implemented as comparisons. Tab. 4 shows the comparison of the experiment results on the literature and the proposed model. The model parameters are shown in Tab. 5. With a slight improvement in the model parameters, the proposed model shows much better performance than the traditional LSTM model. Figs. 2 and 3 present the confusion matrix of the proposed model (LC8). For CNN-based experiments, the audio samples are changed into spectrograms. The spectrograms are used as the input features for the networks.

Tab. 4 Comparison of the experiment results on the literature and the proposed model %

Model	Accuracy	
	eNTERFACE	GEMEP
LC8	90.4	59.0
Ref. [5]	86.2	55.5
Ref. [7]	56.2	36.5
Ref. [8]	89.6	57.0
Ref. [22]	64.2	45.5

Tab. 5 Comparison of the model parameters in the traditional models and proposed models

Model	Backbone	M	Output	Parameter/10^5
L0	LSTM	0		7.18
L8	LSTM	8		7.83
LT8	LSTM	8	$\boldsymbol{O}_{\text{tfl}}$	7.84
LM8	LSTM	8	$\boldsymbol{O}_{\text{mo}}$	7.84
LC8	LSTM	8	$\boldsymbol{O}_{\text{c}}$	8.50

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	92.68	0.00	4.88	0.00	0.00	2.44
Disgust	2.56	89.74	5.13	0.00	0.00	2.56
Fear	0.00	6.98	90.70	0.00	2.33	0.00
Happy	0.00	0.00	0.00	100.00	0.00	0.00
Sad	2.22	0.00	0.00	0.00	97.78	0.00
Surprise	5.26	7.02	7.02	3.51	0.00	77.19

Fig. 2 Confusion matrix of the proposed model on the eNTERFACE corpus

Amusement	81.82	0.00	0.00	9.09	0.00	9.09	0.00	0.00	0.00	0.00	0.00	0.00
Anxiety	13.33	26.67	13.33	13.33	6.67	13.33	0.00	13.33	0.00	0.00	0.00	0.00
Irritation	0.00	4.17	62.50	0.00	4.07	4.17	0.00	4.17	4.17	4.17	4.17	8.33
Despair	0.00	0.00	0.67	33.33	26.67	0.00	0.00	20.00	6.67	0.00	0.00	6.67
Joy	5.00	0.00	15.00	5.00	45.00	5.00	5.00	5.00	10.00	5.00	0.00	0.00
Panic	11.76	0.00	11.76	5.88	5.88	64.71	0.00	0.00	0.00	0.00	0.00	0.00
Anger	0.00	0.00	0.00	15.00	5.00	10.00	70.00	0.00	0.00	0.00	0.00	0.00
Interest	0.00	0.00	10.00	0.00	10.00	0.00	0.00	50.00	20.00	0.00	0.00	10.00
Pleasure	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.67	93.33	0.00	0.00	0.00
Pride	0.00	14.29	0.00	0.00	4.76	0.00	4.76	4.76	9.52	61.90	0.00	0.00
Relief	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9.09	0.00	81.82	9.09
Sadness	0.00	14.29	9.52	0.00	4.76	0.00	0.00	4.76	4.76	0.00	14.29	47.62
	Amusement	Anxiety	Irritation	Despair	Joy	Panic	Anger	Interest	Pleasure	Pride	Relief	Sadness

Fig. 3 Confusion matrix of the proposed model in the GEMEP corpus

Based on the results, the LSTM models show better performance than CNN networks on GEMEP and eNTERFACE corpus. The UARs of the models[7,22] are at least 10% lower than those of the LSTM models. Among all the models, the best UARs (90.4% and 59.0%) are achieved by the proposed model. Therefore, combining the multi-head attention mechanism along with multiple context vectors results in improvement, providing an effective method for SER tasks.

3 Conclusions

1) In this research, an MHTA weighting method is proposed to distinguish the salience regions of emotional speech samples.

2) To form the parts of the input of the full connection layer, the output of the feature-dimension attention layer and last time-step output are utilized. Moreover, feature-dimension attention is employed for the all-time output of the first LSTM layer to screen information for the fully connected layer input.

3) Evaluations are performed on the eNTERFACE and GEMEP corpora. The proposed model achieves the best performance for SER compared with the other models. The results demonstrate the effectiveness of the proposed attention-based LSTM model.

References

[1] Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction[J]. *IEEE Signal Processing Magazine*, 2001, **18**(1): 32 − 80. DOI: 10.1109/79.911197.

[2] Anagnostopoulos C N, Iliou T, Giannoukos I. Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011[J]. *Artificial Intelligence Review*, 2015, **43**(2): 155 − 177. DOI: 10.1007/s10462-012-9368-5.

[3] 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. DOI: 10.1016/j.patcog.2010.09.020.

[4] Trigeorgis G, Ringeval F, Brueckner R, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network[C]//*IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). Shanghai, China, 2016: 5200 − 5204. DOI: 10.1109/ICASSP.2016.7472669.

[5] Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention[C]//*IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). New Orleans, LA, USA, 2017: 2227 − 2231. DOI: 10.1109/ICASSP.2017.7952552.

[6] Tarantino L, Garner P N, Lazaridis A. Self-attention for speech emotion recognition[C]//*Interspeech*. Graz, Austria, 2019: 2578 − 2582. DOI: 10.21437/Interspeech.2019-2822.

[7] Li Y, Zhao T, Kawahara T. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning[C]//*Interspeech*. Graz, Austria, 2019: 2803 − 2807. DOI: 10.21437/Interspeech.2019-2594.

[8] Xie Y, Liang R, Liang Z, et al. Speech emotion classification using attention-based LSTM[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019, **27**(11): 1675 − 1685. DOI: 10.1109/TASLP.

2019.2925934.

[9] Li R, Wu Z, Jia J, et al. Dilated residual network with multi-head self-attention for speech emotion recognition [C]//*IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). Brighton, UK, 2019: 6675 - 6679. DOI: 10.1109/ICASSP.2019.8682154.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//*Advances in Neural Information Processing Systems*. Long Beach, CA, USA, 2017: 5998 - 6008. DOI: 10.7551/mitpress/8385.003.0003.

[11] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//*NAACL-HLT*. Minneapolis, MN, USA, 2019: 4171 - 4186. DOI: 10.18653/v1/n19-1423.

[12] Lian Z, Tao J, Liu B, et al. Unsupervised representation learning with future observation prediction for speech emotion recognition[C]//*Interspeech*. Graz, Austria, 2019: 3840 - 3844. DOI: 10.21437/Interspeech.2019-1582.

[13] Tian Z, Yi J, Tao J, et al. Self-attention transducers for end-to-end speech recognition [C]//*Interspeech*. Graz, Austria, 2019: 4395 - 4399. DOI: 10.21437/Interspeech.2019-2203.

[14] Wang F, Jiang M, Qian C, et al. Residual attention network for image classification [C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, Hawaii, USA, 2017: 3156 - 3164. DOI: 10.1109/CVPR.2017.683.

[15] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, Nevada, USA, 2016: 770 - 778. DOI: 10.1109/CVPR.2016.90.

[16] Eyben F, Wöllmer M, Schuller B. Opensmile: The munich versatile and fast open-source audio feature extractor [C]//*Proceedings of the* 18*th ACM international conference on Multimedia*. New York, USA, 2010: 1459 - 1462. DOI: 10.1145/1873951.1874246.

[17] Schuller B, Steidl S, Batliner A, et al. The INTERSPEECH 2010 paralinguistic challenge[C]//*Interspeech*. Makuhari, Japan, 2010: 2794 - 2797. DOI: 10.21437/Interspeech.2010-739.

[18] Hochreiter S, Schmidhuber J. Long short-term memory [J]. *Neural Computation*, 1997, **9**(8): 1735 - 1780. DOI: 10.1162/neco.1997.9.8.1735.

[19] Martin O, Kotsia I, Macq B, et al. The eNTERFACE05 audio-visual emotion database[C]//*International Conference on Data Engineering Workshops*. Atlanta, GA, USA, 2006: 8. DOI: 10.1109/ICDEW.2006.145.

[20] Bänziger T, Mortillaro M, Scherer K R. Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception[J]. *Emotion*, 2012, **12**(5): 1161 - 1179. DOI: 10.1037/a0025827.

[21] Xie Y, Liang R, Liang Z, et al. Attention-based dense LSTM for speech emotion recognition[J]. *IEICE Transactions on Information and Systems*, 2019, **102**(7): 1426 - 1429. DOI: 10.1587/transinf.2019EDL8019.

[22] Zhao J, Mao X, Chen L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks[J]. *Biomedical Signal Processing and Control*, 2019, **47**: 312 - 323. DOI: 10.1016/j.bspc.2018.08.035.

基于多头注意力长短期记忆模型的语音情感识别方法

赵 焱[1] 赵 力[1] 路 成[1] 李溯南[1] 唐传高[2] 连海伦[1]

([1] 东南大学信息科学与工程学院, 南京 210096)

([2] 东南大学生物科学与医学工程学院, 南京 210096)

摘要:针对语音情感识别中不同表征空间的信息利用不足问题,提出了一种多头注意力的双层长短时记忆模型,用于充分挖掘有效的情感信息.该模型以具有时序情感信息的帧级别特征作为输入值,利用长短时记忆模块学习时域特征,设计了特征注意力模块和时间多头注意力模块,对长短时记忆模块的逐层输出值、特征注意力模块输出值、时间多头注意力模块输出值进行融合.结果表明,相比传统的长短时记忆模型,所提方法在 eENTERFACE 和 GEMEP 两个数据集上的识别准确率分别提升了 14.6% 和 10.5%,从而证明了其在语音情感识别任务中的有效性.

关键词:语音情感识别;长短期记忆;多头注意力机制;帧级别特征;自注意力

中图分类号:TP37