Research of Attention-LSTM Model for Baby Cry Detection Robot

Tianye Jian[1,2], Yizhun Peng [1,2*], Wanlong Peng[1,2], Zhou Yang[1,2]

[1]*College of Electronic Information and Automation, Tianjin University of Science and Technology, Tianjin, 300222, China*

[2]*Advanced Structural Integrity International Joint Research Centre, Tianjin University of Science and Technology, Tianjin, 300222, China*

** pengyizhun@tust.edu.cn*

Abstract

In order to achieve the effective acquisition of frame-level speech features under different emotional needs of baby, a speech emotion recognition model for baby based on improved long-term and short-term memory (LSTM) network is established. The frame-level speech features are used instead of the traditional statistical features to preserve the temporal relationships in the original speech, and the traditional forgetting and input gates are transformed into attention gates by introducing an attention mechanism, in order to improve the performance of speech emotion recognition, the depth attention gate is calculated according to the self-defined depth strategy. The results show that, in Fau Aibo Children's emotional data corpus and baby crying emotional needs database, compared with the traditional LSTM based model, the recall rate and F1 score of this model are 3.14% , 5.50% , 1.84% and 5.49% higher, respectively, compared with the traditional model based on Lstm and GRU, the training time is shorter and the speech emotion recognition rate of baby is higher.

Key words: children's emotional; time sequence relationship; frame-level speech feature; deep attention gate; Long Short-Term Memory(LSTM)

1. Introduction

Children's emotion recognition is an important part of affective computing[1]. Children are far less able than adults to act rationally in emotional outbursts and in response to different emotions, which can lead to emotional disorders if children are not able to act rationally and are directed in a timely manner, which can lead to anxiety disorders and other mental health problems. Therefore, it is of great significance to use appropriate algorithms or models to judge and guide children's emotions.

The researchers conducted in-depth research on children's emotion recognition from acoustic features, machine learning and deep learning[2] it is proposed to use Support vector machine and convolutional neural network to construct a system for detecting children's secondary emotional states[3] real-time emotional state of children is defined by multi-agent based interaction system[4] to establish the children's dual-modal emotion database and use the dual-modal emotion recognition method to measure the proportion of children's emotion contribution, and to point out that infants'(or young children's) emotion is more difficult to judge than older children's, babies usually cry to express their needs to their parents or guardians[5] the Mel-Frequency Cepstrum Coefficients (MFCC) of infant cries were extracted and classified based on Hidden Markov Model (Hmm) to identify whether the infant cries were healthy or not6. The Spectrogram was used as the feature vector, and the Convolutional Neural Network (CNN) was selected as the classification model to classify and recognize the crying of infants in pain, hunger and sleepiness[7]. The Support Vector Machine (SVM) was used as the classifier to classify the crying sounds of infants in the condition of hunger, pain and sleepiness, and the recognition effect was better than that of the Support Vector Machine (SVM).

Although the above algorithms have been successfully applied to children's emotion recognition, traditional machine learning algorithms, as well as self encoders and convolutional neural network in deep learning, can only accept data with fixed dimensions as input, this is in contradiction with the fact that the effective length of speech is constantly changing. To solve this problem, reference extracts emotion-related features (hereinafter

referred to as frame-level features) from short-term speech frames, and applies static statistical functions (such as mean, variance, maximum, linear regression coefficient) to frame-level features, finally, the feature vectors with fixed dimensions are formed to represent the features of the frame speech[8-10]. Although this method solves the problem of model input, the time sequence information of the original speech is lost through the statistical analysis of the speech features.

2. Related work

2.1 LSTM network

The LSTM Network is a variant of Recurrent Neural Network (RNN), which is mainly used to process sequence information with long time difference. The LSTM network can solve the problem that the long-term information is difficult to store because of the gradient disappearance of RNN in the reverse propagation[11-13]. The LSTM network has been successfully applied in natural language processing (NLP) [14-16]. In order to enhance the ability of LSTM network to process data in specific tasks, the researchers further optimized the internal structure of LSTM network[17]. The fusion of the input gate and the forgetting gate of the LSTM network by the Gated Recurrent Unit (GRU) reduces the model parameters, the performance of LSTM network is better than that of GRU[18] in all machine translation tasks[19]. By using CONVLSTM network structure, the computing method of gate structure of LSTM is improved from Matrix multiplication to convolution[20]. Infinite Impulse Response Filter (IIR) Memory block of RNN is improved to Finite Impulse Response Filter (Fir) Memory block by Feedforward Sequential Memory Network (FSMN)[21], but FSMN usually needs to stack very deep layers, so FSMN has delay compared with one-way LSTM network[22]. Advanced short-term memory (Advanced LSTM) network, which uses attention mechanism to weight multiple cell states, can be effectively used for emotion recognition. However, reference[23] indicates that this method does not change the gate structure in the LSTM network and requires more training time. In addition, researchers are exploring how to stack LSTM structures to achieve more reliable emotion recognition[24]. End-to-end emotion recognition is achieved by extracting multi-channel speech features from a 6s-long speech waveform via a Convolutional Neural Network (CNN) as input to the LSTM Network[25]. In this paper, 1280 kinds of abstract features were extracted from 6s-long speech waveform by CNN, and then fused with facial features as input of LSTM network.

In this paper, we propose an improved Long Short-Term Memory (LSTM) based model for children's speech emotion recognition. Based on the LSTM network structure, the frame-level speech features are substituted for the traditional statistical features, in order to obtain better recognition performance, attention gates were used to replace the traditional forgetting gates and input gates, and to construct the deep attention gates by weighting attention in multiple cellular states.

The formula used in the traditional LSTM network is as follows:

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

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

$$C_t = tanh(w_C . [h_{t-1}, x_t] + b_c \quad (3)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t \quad (4)$$

$$O_t = \sigma(w_o . [h_{t-1}, x_t] + b_o \quad (5)$$

$$h_t = o_t . \tanh(C_t) \quad (6)$$

σ Is the sigmoid activation function, $h_{t-1}$is the hidden layer output at $t-1$, x_t is the input at t, C_{t-1} is the cell state at $t-1$, $C_t$is the candidate value of the cell state at t, f_t、i_t、o_t. i_t and o_t are respectively forgetting gate, input gate and output gate.

2.2 Attention Mechanisms

The attention mechanism is based on the human visual attention mechanism. Attention makes people pay more attention to the important part of the information captured by vision, get as much detail information as possible, and reduce the attention to the irrelevant information around the target, that is, to suppress the irrelevant information[26]. In order to effectively utilize the information output from the LSTM network, the soft attention mechanism is introduced into the LSTM Network Model (hereinafter referred to as the LSTM model), and it is successfully applied to the field of machine translation, by weighting the attention of the LSTM model at different time, the relevance degree of the word to be translated and other words can be

expressed. The encoding-decoder structure based on attention mechanism is proposed in reference[27], and its application in speech recognition is superior to HMM decoding system. Based on the encoder-decoder structure, a local attention model is proposed in reference [28], in which an alignment position is first predicted, and then a probability distribution similar to the soft attention model is obtained over the l window at the alignment position[29]. The single-head attention mechanism is improved to the multi-head attention mechanism, and the quality of machine translation is improved significantly by Transformer model.

In recent years, the mechanism of self-attention has become a hot topic[30]. Calculates the self-attention of the output of the LSTM model, calculates several fractions for different time steps, and then proposes a new LSTM model:

$$A = softmax(W_{s2} \cdot tanh(W_{s1} \times H^T)) \quad (7)$$

Where a is the attention fraction matrix, hi is the output of hidden layer cells at time I, and h is the stack result of hi in each time of LSTM MODEL:

$$H = (h_1, h_2, \dots, h_n) \quad (8)$$

The weighted output is expressed as:

$$M = A \times H \quad (9)$$

The introduction of the attention mechanism reduces the computational burden of processing Galway input data, and makes the task processing system more focused on finding information in the input data that is significantly related to the current output, thus improving the output quality. In recent years, researchers have applied attention mechanisms to the improvement of affective effects in speech[31]. The attention mechanism was used to screen the features and cross-link between the multilayer LSTM networks, and good effect of emotion recognition was obtained. In the output of RNN, a Local Attention mechanism is proposed to improve the effect of emotion recognition in multiple data sets[32].

3. Improved Deep Attention Gate

3.1 Attention Gate

In this paper, the attention mechanism is introduced into the internal structure of LSTM model, and the attention gate-based LSTM model is proposed, which greatly reduces the number of parameters of LSTM By introducing the concept of depth into the attention gate, the LSTM model can learn the input features better and avoid information redundancy. The attention-gate-based LSTM structure proposed in this paper enables the cell state at the last moment to determine the characteristics to be noticed when calculating each time step, at the same time, the attention gate is used to modify the traditional forgetting gate and the input gate is used to weight the features that need attention.

Because traditional input gates and forgetting gates are only implemented by a single fully connected layer, the model needs to be trained enough times to notice the cellular state information that needs to be left and the new input information that needs to be added, which causes it to converge and slow down[33]. On this basis, the peephole connection is added, the cell state is also taken as input, the cell state information is added into the three gates, and the increase of the parameters results in the increase of the training time and space complexity. In this paper, self-attention is paid to every cell state, and input candidate information is added to the parts of the cell state that do not need attention, the self-attention Algorithm replaces the three matrices needed for forgetting, input, and peeping connections with attention gates.

The attention gate at is defined as follows:

$$a_t = activation(V \cdot tanh(W \cdot C_{t-1})) \quad (10)$$

The activation is an activation function (the activation function can be selected according to the need, but its value range should be less than 1). The calculation formula for updating the Cell State is:

$$C_t = a_t \cdot C_{t-1} + (1 - a_t) \cdot C_t \quad (11)$$

Attention Gate can improve the recognition rate of the model while reducing the number of parameters and training time. Model distillation[34], 8-bit quantization[35], and shared parameter[36-37] are usually used in the existing reports. In this paper, an attention gate based on attention mechanism is proposed, which significantly reduces the parameters of LSTM model. In addition, the attention-gate-based model can reduce

more training time for longer input sequences due to the modification within the LSTM model. For example, for a layer LSTM model with an input dimension of 512 and an output dimension of 256, if the bias is ignored, the commonly required parameters are: 1) the dimension weight of [512 + 256,256 × 4] for the three gate structures and the candidate values; 2) if the cell state at the last moment is taken into account in calculating the gate structure, a vector of 3 × [1,256] is added as the peephole[38]. In this paper, we do not need to introduce the peephole structure because we calculate the self-attention as the attention gate directly to the cell state. At the same time, the number of parameters required is reduced to [512 + 256,256 × 2] and the weight of 2 × [256,256] for calculating attention due to the combination of forgetting and input gates as attention gates. For this layer, the number of parameters was reduced from the original 787200 to 524288, reducing the number of parameters by 33.4%. For the LSTM model with deeper layers, more complex model and more data, the parameter quantity is reduced effectively.

3.2 Deep Attention Gate

The LSTM model is usually used to process time series information, but this information will increase with time, therefore, the calculation of the LSTM model at a certain time (that is, updating the cell state C and hidden layer output h) is only based on the external input and the cell state and hidden layer output at the previous time. Before the attention mechanism is put forward, if each moment is considered several times before, it will lead to too much information and the loss of important information, and increase the calculation amount and lead to the gradient explosion. However, the information of cell state at t-time is not only related to the information at t-1, but also closely related to the information at t-2, which is selectively forgotten at t-1. In this paper, the concept of deep forgetting gate is proposed and the corresponding input gate is designed.

The deep forgetting gate looks not only at the information about the state of the cell at the last moment (depth length = 1) , but also at the information about the state of the cell at the time t-2, t-3, ... , t-n (depth length = n) , which constructs the Deep-Attention-Lstm Cell structure, as shown in Fig.1

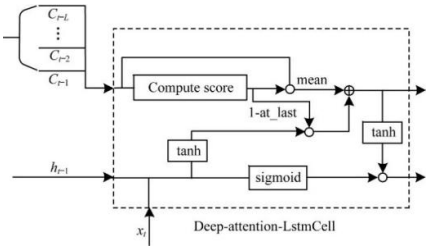

Fig.1 Deep-Attention-Lstm Cell internal structure sketch

It is worth noting that the introduction of “depth” will lead to an increase in training time. This is because in addition to forward increasing the number of attention gates in the loop for computing the state of multiple cells, reverse propagation also increases the number of chain derivations. From the point of view of the parameters of the model, although the depth will cause the increase of the training time, it will not cause the increase of the parameters of the model because the attention gate weights of each layer are shared by V and W.

In this paper, the aim of depth is to improve the performance of speech emotion recognition. In order to study the performance improvement, the following experiments were carried out:

In experiment 1 investigated the effect of depth performance on children’s emotion recognition rate. The attention-gate-based LSTM model (hereafter referred to as the attention-gate LSTM model) at Depths 1, 2, and 3 is used for comparison.

In experiment 2, the effects of decreasing the number of parameters and training time on speech emotion recognition performance were investigated. The attention gate LSTM model with depth 1 was compared with the traditional GRU model and LSTM model.

3.3 Training Framework

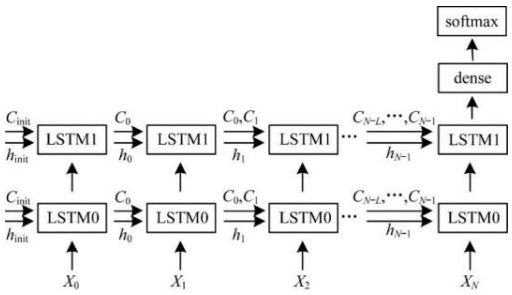

Fig. 2 Training framework of depth of attention gate LSTM model

The training framework for the deep attention gate LSTM model is shown in Fig.2. Among them, LSTM0 represents the LSTM model of the first-level deep attentional gate, and LSTM1 represents the LSTM model of the second-level deep attentional gate. X_t extracts speech features [8-10] from INTERSPEECH in frame t after framing and windowing, h_t and C_t are the hidden layer output and cell state of its corresponding LSTM model output. As can be seen from Fig. 2, the input state of the traditional LSTM model at time t is (h_{t-1}, C_{t-1}) , while in the training of this paper, the input state at each time is expanded to (h_{t-1}, { C_{t-1}, C_{t-2}, ... , C_{t-L} }) , where L is the depth of the attention gate. The last state of the last layer of LSTM, which contains all the temporal information of the preorder, is input into the subsequent classification network for the identification of children's emotions.

4. Experimental Setup and Analysis

4.1 Experimental Setup

In the experiment, two databases with different emotional representation forms are used to verify the validity of the algorithm in this paper. In order to study the performance of this algorithm in dealing with other types of emotion recognition problems, and whether reducing the number of parameters can optimize the time or reduce the performance, Fau Aibo Children's affective Corpus, infant crying affective needs Corpus and CASIA Chinese affective Corpus were used to verify the results[39].

FAU AIBO EMOTIONAL Corpus for children: Fau Aibo used high performance wireless earphones to collect and record the vocalization of 51 children aged around 10 years old and their electronic pet Aibo during the game, with the most emotionally salient data retained. Natural language contains 48,401 words. In order to ensure the accuracy of the tagging, each sample of the Corpus is auditioned by 5 language majors and then marked by voting. This paper selects five categories of tags defined in the INTERSPEECH 2009 emotional challenge: A (Angry, Touchy, Reprimanding) , E(Emphatic) , N (Neutral) , P(Motherese、Joyful) and R (Rest) .

2) Infant crying emotional needs corpus: Since there is no universal crying emotional needs Corpus in the world, the author collaborated with a hospital in China to record and annotate the audio files of infant crying in five states: Angry, Hungry, Pain, Sad and Tired. In order to improve the quality of the Corpus, the author screened the emotional speech data of infant crying by artificial method, and got rid of the speech-related frames of parents comforting children, and the speech-related frames of two or more babies crying at the same time. The Corpus consists of 10 infants (5 boys and 5 girls respectively), each infant has 20 items in each state, which is 5 × 10 × 20 = 1000 items.

3) CASIA Chinese affective Corpus was recorded by Chinese Academy of Sciences, and the pronunciations of six emotions, namely angry, happy, fear, sad, surprise and neutral, were carried out by four related professionals. There are 9600 Corpora in this corpus.

4.2 Frame Level Feature Selection

The experiment selects part of the frame-level features based on the speech emotion features of INTERSPEECH. In reference [8], 16 low-level descriptors (LLD, zero-crossing rate, root-mean-square Frame Energy, pitch frequency and Mel Frequency Cepstrum Coefficients 1-12) and their difference coefficients are extracted. For each descriptor, 12 statistical functions are calculated, so the total eigenvector has 16 × 2 × 12 = 384 features. On this basis, the speech emotion characteristics of INTERSPEECH 2010(IS2010) have been increased to 38 kinds of LLD, so the total feature dimension has been expanded to 1582 dimensions. The feature dimension of the speech feature set is increased to 6373 dimensions.

4.3 Setting of Experimental Parameters

The original data is divided into two parts, the training set and the test set, which are separated from each other, and the ratio of the training set and the test set is 4:1. The experiments all adopt unidirectional two-layer LSTM stack structure and use one full connection layer and one $softmax$ layer as the training model. During the training, a small batch of gradient descent is used and $Tanh$ is used as the activation function, as shown in Table 1. In order to ensure the validity of experimental comparison, the same Corpus and model experimental parameters are identical.

5. Algorithm Performance Analysis

The traditional LSTM model gets rid of the redundant information through the forgotten gate and gets the new information through the input gate. In this paper, we use the self-attention and the basic structure of LSTM to do self-attention to the cell state, so as to compare the forgetting gate and the input gate of LSTM. At the same time, considering the correlation of time series information, the depth-based self-attention Gates are proposed and compared at the depth of 1, 2 and 3. Four kinds of models were compared: Traditional model, LSTM + deepf_1 model, LSTM + deepf_2 model, and LSTM + deepf_3 model. The depth distributions of these models are 0, 1, 2 and 3, as shown in Fig.3. As can be seen from Fig.3 and Fig.4 , after replacing the forgetting door and the output door of the traditional

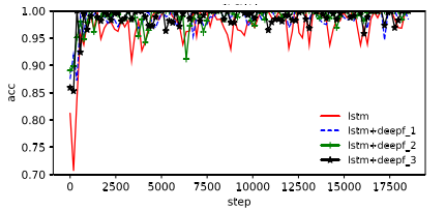

Fig. 3The performance of infant crying emotional needs corpus in different LSTM models

LSTM model with the attention door proposed by using

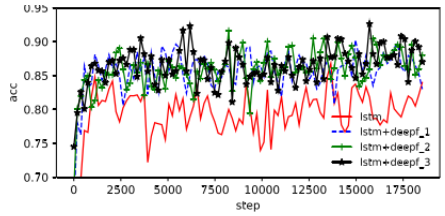

Fig. 4 Performance of different LSTM models using Fau Aibo children's emotion corpus database

the infant crying affective needs corpus and Fau Aibo children affective Corpus, the convergence rate of the attention gate LSTM model on the training set and the test set is much higher than that of the traditional LSTM model, when FAU AIBO is used in children's affective Corpus, the traditional LSTM model starts to converge at about 30000, while the attention gate model starts to converge at about 17000 When the model converges, the attention-gate LSTM model outperforms the traditional LSTM model in the average recognition rate

Table 1. Experimental parameters.

Name of Parameter	Baby crying needs corpus parameter values	Fau Aibo	CASIA
Eta	1e-3 beta2= 0.7	1e-4 beta2= 0.9	1e-4 beta2= 0.9
Batch size	64	128	128
Epochs	1 500	1 200	1 200
Lstm cells	[512, 256]	[512, 256]	[512, 256]
Dense layers	[256, 128]	[256, 128]	[256, 128]
Softmax layers	[128, 5]	[128, 5]	[128, 6]
L2	1e-4	1e-4	1e-4

of children's emotion.

According to the analysis above, the performance of attention gate LSTM model is improved because it modifies the forgetting gate and input gate of traditional LSTM model, it makes the LSTM model leave important information by self-attention to the cell state at the last moment, and supplement the unimportant information as the new input in the corresponding position, thus improving the performance of the LSTM model. The attention gate LSTM model introduces the concept of depth so that each forgetting operation is determined by multiple cell states rather than one of them.

In order to compare the performance of different models in the test set for each kind of emotion, the performance index of the model with the highest recognition rate from the beginning of training to the end of the test set was Quantitative analysis, the performance indicators

obtained from the infant crying affective needs Corpus and Fau Aibo children affective corpus are shown in tables 2 and 3. It can be seen that for the test set, the performance index of the attention gate LSTM model is better than the traditional LSTM model.

From Table 2, it can be seen that the recall rate of attention-gate LSTM model is better than that of traditional LSTM model except that the items of drowsiness and sleepiness are close to each other The F1 scores of attention gate LSTM model were better than those of traditional LSTM model in five kinds of emotion. In the aspect of depth, the performance of the attention-gate LSTM model of depth 3 and depth 2 is close to that of the attention-gate LSTM model of depth 1, except "sad", the recall rate and F1 score of the other 4 items of the model are better than those of depth 1.

From Table 3, it can be seen that when FAU AIBO is used in children's affective corpus, the recall rate and F 1 score of attention gate LSTM model are lower than those of traditional LSTM model except that the e class is lower than that of traditional LSTM model, and the other four items are better than traditional LSTM model. In the aspect of depth, the performance of the attentional gate LSTM model of depth 3 and depth 2 is close to that of the attentional gate LSTM model of depth 1, except for class R, the recall rate and F1 score of the other four items of the attentional gate LSTM model are better than those of depth 1.

Table 2. Performance indicators of different LSTM models using emotional needs corpus of infant crying database.

model	Measure	Angry	Hungry	Pain	Sad	Tired	AVG
sample size	—	40	45	34	35	46	200
LSTM	Recall	0.875	0.911	0.824	0.829	0.957	0.885
	F1 SCORE	0.875	0.901	0.848	0.906	0.889	0.885
LSTM+deepf_1	Recall	0.875	0.889	0.882	1.000	0.935	0.915
	F1 SCORE	0.909	0.941	0.822	0.946	0.945	0.916
LSTM+deepf_2	Recall	0.925	0.978	0.882	0.943	0.957	0.940
	F1 SCORE	0.949	0.957	0.923	0.930	0.936	0.940
LSTM+deepf_3	Recall	0.925	0.978	0.824	0.943	0.957	0.930
	F1 SCORE	0.949	0.926	0.875	0.943	0.946	0.930

Table 3. Performance indicators of different LSTM models using Fau Aibo children's emotion corpus database.

model	Measure	A	E	N	P	R	AVG
sample size	—	611	1 508	5 376	215	546	8 256
LSTM	Recall	0.352	0.373	0.755	0.088	0.081	0.594
	F1 SCORE	0.339	0.373	0.743	0.110	0.093	0.586
LSTM+deep f_1	Recall	0.326	0.300	0.814	0.153	0.119	0.621
	F1 SCORE	0.358	0.338	0.770	0.173	0.133	0.603
LSTM+deep f_2	Recall	0.339	0.289	0.826	0.158	0.081	0.625
	F1 SCORE	0.342	0.340	0.772	0.173	0.104	0.601
LSTM+deep f_3	Recall	0.360	0.327	0.800	0.191	0.095	0.619
	F1 SCORE	0.371	0.363	0.765	0.192	0.112	0.604

It should be noted that the sample size of Fau Aibo Children's affective corpus is uneven, with a maximum of 5,376 samples in n category and only 215 samples in P category. From the above analysis, with the increase of depth, the model can enhance the learning of a small number of samples. Compared with the traditional LSTM model, the recall rate of LSTM + deepf_1model was increased by 5.50% , and the F 1 score was increased by 5.49% , and the recall rate of LSTM + deepf_2 model was increased by 3.14% when Fau Aibo was used, the F1 score of LSTM + deepf_3 model was increased by 1.84% .

6. Conclusion

In this paper, a child speech emotion recognition model based on improved LSTM network is proposed frame-level speech features are used instead of traditional speech features, the attention mechanism is introduced into the forgetting gate and the input gate of the internal structure of the LSTM network model to form the attention gate. The experimental results show that the recognition rate of this model is significantly higher than that of traditional LSTM model, and the recognition rate of depth model is higher than that of shallow model. In CASIA database with other emotions, the training time of this model is shorter than that of LSTM model, and the recognition rate is higher than that of LSTM model and GRU model. The next step is to introduce this model into the fields of speech recognition, machine translation and lie detection, to test and study the continuous affective Corpus and to improve the model for calculating attention scores, to further improve children's speech emotion recognition rate.

References

1. LU Fang, CHEN Guopeng. An overview on the development of children's emotion regulation[J]. Psychological Science, 2003, 26(5): 928-929. (in Chinese)
2. GONG Y, POELLABAUER C.Continuous assessment of children's emotional states using acoustic analysis[C]//Proceedings of 2017 IEEE International Conference on Healthcare Informatics.Park City, USA: IEEE Press, 2017: 171-178.
3. DE SILVA P R, MADURAPPERUMA A P, MARASINGHE A, et al.A multi-agent based interactive system towards Child 's emotion performances quantified through affective body gestures[C]//Proceedings of the 18th International Conference on Pattern Recognition.Hong Kong, China: IEEE Press, 2006: 1236-1239.
4. DAI Weijia.Research on expression and speech bimodal emotion recognition of children[D].Nanjing: Southeast University, 2016.
5. LEDERMAN D, COHEN A, ZMORA E, et al.On the use of hidden Markov models in infants'cry classification[C]//Proceedings of the 22nd Convention on Electrical and Electronics Engineers in Israel.Tel-Aviv, Israel: IEEE Rress, 2002: 350-352.
6. CHANG Chuanyu, LI Jiajing.Application of deep learning for recognizing infant cries[C]//Proceedings of 2016 IEEE International Conference on Consumer Electronics-Taiwan.Nantou County, China: IEEE Press, 2016: 1-2.
7. CHANG C Y, CHANG C W, KATHIRAVAN S, et al. DAG-SVM based infant cry classification system using sequential forward floating feature selection[J]. Multidimensional Systems and Signal Processing, 2017,

28(3): 961-976.

8. SCHULLER B, STEIDL S, BATLINER A.Theinterspeech 2009 emotion challenge[C]//Proceedings of the 10th Annual Conference of the International Speech Communication Association 2009.Brighton, UK: IEEE Press, 2009: 1-7.
9. SCHULLER B, STEIDL S, BATLINER A, et al.The INTERSPEECH 2010 paralinguistic challenge[C]//Proceedings of the 11th Annual Conference of the Inter-national Speech Communication Association 2010.Makuhari, Japan: IEEE Press, 2010: 25-32.
10. SCHULLER B, STEIDL S, BATLINER A, et al.The INTERSPEECH 2016 computational paralinguistics challenge: deception, sincerity and native language[C]//Proceedings of Interspeech 2016.San Francisco, USA: ISCA, 2016: 2001-2005.
11. HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
12. HAN Wenjing, LI Haifeng. A brief review on emotional speech databases[J]. Intelligent Computer and Applications, 2013, 3(1): 5-7.
13. ATHIWARATKUN B, STOKES J W.Malware classification with LSTM and GRU language models and a character-level CNN[C]//Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing.New Orleans, USA: IEEE Press, 2017: 2482-2486.
14. MERITY S, KESKAR N S, SOCHER R.Regularizing and optimizing LSTM language models[EB/OL].(2017-08-07)[2019-10-20].
15. LI Wei, MAK B.Derivation of document vectors from adaptation of LSTM language model[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics.Stroudsburg, USA: Association for Computational Linguistics, 2017: 456-461.
16. LIU Yue, ZHAI Donghai, REN Qingning. News text classification based on CNLSTM model with attention mechanism[J]. Computer Engineering, 2019, 45(7): 303-308.
17. CHO K, VAN M B, GULCEHRE C, et al.Learning phrase representations using RNN encoder-decoder for statistical machine translation[EB/OL].(2014-06-03)[2019-10-20].
18. BRITZ D, GOLDIE A, LUONG M T, et al.Massive exploration of neural machine translation architectures[EB/OL].(2017-03-11)[2019-10-20].
19. SHI Xingjian, CHEN Zhourong, WANG Hao, et al.Convolutional LSTM network: a machine learning approach for precipitation nowcasting[EB/OL].(2015-06-13)[2019-10-20].
20. ZHANG Shiliang, JIANG Hui, WEI Si, et al.Feedforward sequential memory neural networks without recurrent feedback[EB/OL].(2015-10-09)[2019-10-20].
21. ZHANG Shiliang, LEI Ming, YAN Zhijie, et al.Deep-FSMN for large vocabulary continuous speech recognition[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing.Calgary, Canada: IEEE Press, 2018: 5869-5873.
22. TAO Fei, LIU Gang.Advanced LSTM: a study about better time dependency modeling in emotion recognition[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing.Calgary, Canada: IEEE Press, 2018: 2906-2910.
23. XIE Yue, LIANG Ruiyu, LIANG Zhenlin, et al. Speech emotion classification using attention-based LSTM[J]. ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(11): 1675-1685.
24.. TRIGEORGIS G, RINGEVAL F, BRUECKNER R, et al.Adieu features end-to-end speech emotion recognition using a deep convolutional recurrent network[C]//Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing.Shanghai, China: IEEE Press, 2016: 5200-5204.
25. TZIRAKIS P, TRIGEORGIS G, NICOLAOU M A, et al. End-to-end multimodal emotion recognition using deep neural networks[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1301-1309.
26. BAHDANAU D, HYUN CHO K, BENGIO Y.Neural machine translation by jointly learning to align and translate[EB/OL].(2014-09-01)[2019-10-20].
27. CHOROWSKI J, BAHDANAU D, CHO K, et al.End-to-end continuous speech recognition using attention-based recurrent NN: first results[EB/OL].(2014-12-04)[2019-10-20].
28. LUONG M T, PHAM H, MANNING C D.Effective approaches to attention-based neural machine translation[EB/OL].(2015-08-17)[2019-10-20].
29. VASWANI A, SHAZEER N, PARMAR N, et al.Attention is all you need[C]//Proceedings of Advances in neural information processing systems 2017.Long Beach, USA: IEEE Press, 2017: 5998-6008.
30. LIN Z H, FENG M W, SANTOS C N, et al.A structured self-attentive sentence embedding[EB/OL].(2017-03-09)[2019-10-20].
31. XIE Yue, LIANG Ruiyu, LIANG Zhenlin, et al.

Attention-based dense LSTM for speech emotion recognition[J]. ICE Transactions on Information and Systems, 2019, 102(7): 1426-1429.

32. MIRSAMADI S, BARSOUM E, ZHANG C.Automatic speech emotion recognition using recurrent neural networks with local attention[C]//Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing.New Orleans, USA: IEEE Press, 2017: 2227-2231.
33. GERS F A, SCHMIDHUBER J.Recurrent nets that time and count[C]//Proceedings of Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks.Como, Italy: IEEE Press, 2000: 189-194.
34. HINTON G, VINYALS O, DEAN J.Distilling the knowledge in a neural network[EB/OL].(2015-03-09)[2019-10-20].
35. JACOB B, KLIGYS S, CHEN B, et al.Quantization and training of neural networks for efficient integer-arithmetic-only inference[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.Salt Lake City, USA: IEEE Press, 2018: 1-5.
36. HAN S, MAO H Z, DALLY W J.Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding[EB/OL].(2015-10-01)[2019-10-20].
37. DAI Dawei, YU Liping, WEI Hui. Parameters sharing in residual neural networks[J]. Neural Processing Letters, 2020, 51(2): 1393-1410.
38. HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
39. PAN Shifeng, TAO Jianhua, LI Ya.The CASIA audio emotion recognition method for audio/visual emotion challenge 2011[C]//Proceedings of Affective Computing and Intelligent Interaction.Berlin, Germany: Springer Berlin Heidelberg, 2011: 388-395.
40. JAITLY N, HINTON G.Learning a better representation of speech soundwaves using restricted boltzmann machines[C]//Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing.Prague, Czech Republic: IEEE Press, 2011: 5884-5887.