LETTER

Attention-Based Dense LSTM for Speech Emotion Recognition

Yue XIE[†], ***Nonmember***, **Ruiyu LIANG**[††], ***Member***, **Zhenlin LIANG**[†], ***and*** **Li ZHAO**[†a)], ***Nonmembers***

SUMMARY Despite the widespread use of deep learning for speech emotion recognition, they are severely restricted due to the information loss in the high layer of deep neural networks, as well as the degradation problem. In order to efficiently utilize information and solve degradation, attention-based dense long short-term memory (LSTM) is proposed for speech emotion recognition. LSTM networks with the ability to process time series such as speech are constructed into which attention-based dense connections are introduced. That means the weight coefficients are added to skip-connections of each layer to distinguish the difference of the emotional information between layers and avoid the interference of redundant information from the bottom layer to the effective information from the top layer. The experiments demonstrate that proposed method improves the recognition performance by 12% and 7% on eNTERFACE and IEMOCAP corpus respectively.

key words: *attention mechanism, speech emotion recognition, dense connections, LSTM*

1. Introduction

Speech is one of the primary faucets for expressing emotions, and thus for a natural human-machine interface, it is important to recognize, interpret, and respond to the emotions expressed in speech [1]. Deep learning, as a hot research topic, further promotes the research of speech emotion recognition.

In order to enable the model could make use of temporal information of speech, LSTM network becomes the first choice. LSTM network has been successfully applied in the research of speech emotion recognition. Wollmer [2] first applied LSTM to continuous emotion recognition and extracted 4843 features for each utterance as the input of LSTM. In order to enhance the features, the time window is fed frame by frame into a recurrent layer in [3], and the experiments on emotion classification got a better result. However, the above work does not delve into the internal relationship between LSTM output and emotional information, and usually takes the output at the last moment as the final output, while the emotional saturation of speech is not equivalent in different periods, especially, the silent segments of speech contains less emotional information. Therefore, the attention mechanism is applied into the output of LSTM to distinguish the difference of information.

The attention mechanism was first successfully applied in the field of image processing [4], [5], which is inspired by the fact people tend to pay attention to parts of an image rather than the whole one. The attention mechanism had been applied into LSTM networks. Lin [6] calculated the weight coefficient for each time step of LSTM by the attention mechanism. Luong [7] provided three methods of attention score for the auto-encoder constructed from LSTM. On this basis, two methods of attention weighting is proposed for the LSTMs output of each layer from the time dimension and the feature dimension, which is used for the skip-connections of dense layer to distinguish the diffidence between the outputs of each layer. This algorithm could avoid the interference of the redundant information in the output of the bottom layer to the effective information in the upper layer. In the term of speech-based emotion recognition task, the weighting on the time dimension reflects the difference of emotion saturation among periods, while that on the feature dimension reflects the distinguishability of different features.

2. Attention-Based Skip Connections

The traditional deep learning method often suffers from degradation because of the information loss during back propagation. To some extent, residual neural network (ResNet) [8] had solved this problem by connecting directly the previous layer with the upper layer.

$$F(x) = H(x) - x \tag{1}$$

where x is the input and $H(x)$ is the expected output. When the input is directly transferred to the output as the initial result, then the new learning object becomes $F(x)$. New network only needs to learn the difference between input and output, which simplifies the learning difficulty. However, the components of input matrix x have different contributions to the final task. For example, the impact of different speech features on emotion recognition is different. Therefore, weight is added to the input with skip connection in this study, and the expected output is modified as shown in Eq. (2).

$$H(x) = W_{a\_F}F(x) + W_{a\_x} \bullet x \tag{2}$$

where W_{a_F} and W_{a_x} are weighting parameters, which are obtained by the attention mechanism to distinguish the different components of the input information and avoid the

Manuscript received January 22, 2019.
Manuscript revised March 19, 2019.
Manuscript publicized April 17, 2019.

[†] The authors are with School of Information Science and Engineering, Southeast University, Nanjing, 210096, P.R. China.

[††] The author is with School of Communication Engineering, Nanjing Institute of Technology, Nanjing, 211167, P.R. China.

a) E-mail: zhaoli@seu.edu.cn

DOI: 10.1587/transinf.2019EDL8019

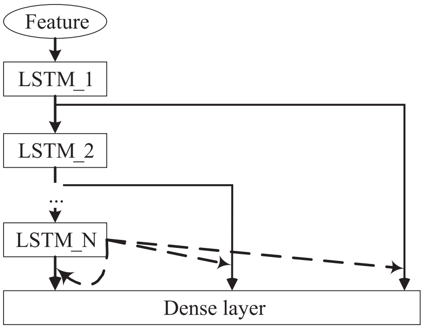

Fig. 1 Structure of model

interference of the redundant information in the output of the bottom layer to the effective information in the upper layer.

The following part gives a detailed introduction to the calculation of weighting parameters in combination with the speech emotion recognition task and the network structure of this study. As shown in Fig. 1, the model structure references skip connections in Dense Convolutional Network [8], [9]. What is difference is that convolutional neural network is replaced with long short-term memory (LSTM) because speech is a typical time series data that is more suitable as the input of LSTM with processing sequence capability. The dotted lines in the figure represent the attention weighting operation that can be obtained in two ways.

2.1 Attention Weighting on Time Dimension

Speech signals contain many silent segments that have less emotional information that means the emotional saturation is different between periods. Therefore, weights can be added to the time dimension of the LSTMs output to distinguish the difference between outputs of layers. In order to achieve the attention weighting of time dimension, this study references the algorithm of Luong [7] who believes that the current target word to be translated depends on components of the source sentence to different degrees. It calculates the global attentional weight score through the target hidden state h_t of the decoder (corresponding to the target word to be translated) and the source hidden state $\bar{h}_s$ of the encoder (corresponding to the original statement).

$$a_t(s) = \mathrm{align}(h_t, \bar{h}_s) = \frac{\exp(\mathrm{score}(h_t, \bar{h}_s))}{\sum_s \exp(\mathrm{score}(h_t, \bar{h}_s))} \tag{3}$$

$$\begin{aligned} &\mathrm{score}_{\mathrm{dot}}(h_t, \bar{h}_s) = h_t^{\mathrm{T}} \bar{h} \\ &\mathrm{score}_{general}(h_t, \bar{h}_s) = h_t^{\mathrm{T}} W_a \bar{h}_s \\ &\mathrm{score}_{concat}(h_t, \bar{h}_s) = v_a^{\mathrm{T}} \tanh(W_a[h_t; \bar{h}_s]) \end{aligned} \tag{4}$$

where v_a and W_a are the trainable parameters. The concat method is the additive attention [10], while the general method is multiplicative attention [7]. Additive attention and multiplicative attention are similar in complexity, but the multiplicative attention is more efficient in storage, because matrix operations can be implemented more efficiently.

Attention weighting on the time dimension is based on the multiplicative attention and takes the characters of LSTMs output into consideration.

$$W_{t\_i} = \mathrm{softmax}(h_{N,last} W_i h_i^T) \tag{5}$$

where softmax function is the Eq. (3), $h_{N,last}$ is the output of the last time step of the last layer of LSTM network (N represents the number of layers), which contains the most effective information because of the memory capacity of LSTM, so it can be used as a parameter to measure the information content of each layers output. h_i represents the output of LSTM in the i-th layer, whose row corresponds to the time information and column corresponds to the feature information. W_{a_i} obtained by the transpose item h_i^T is the score of each time step, which reflects the difference in emotion saturation. In this study, it is denoted as time attention. W_i is the trainable parameter of attention mechanism in the i-th layer.

2.2 Attention Weighting on Feature Dimension

In emotion recognition task, different speech features have different ability to distinguish emotion categories. Therefore, the weighting coefficients can be obtained according to the information differences at the feature level.

$$W_{f\_i} = \mathrm{softmax}(V_i tanh(W_i h_i)) \tag{6}$$

where V_i and W_i are the training parameters. Equation (6) references the self-attention algorithm [6]. The weight of each layer is calculated according to its own output h_i.

The input of the dense layer will be modified according the Eq. (5) and Eq. (6).

$$\begin{aligned} D_{general} &= [h_{1,last}, h_{2,last}, \ldots, h_{N,last}] \\ D_{time} &= [W_{t\_1} h_1, W_{t\_2} h_2, \ldots, W_{t\_N} h_N] \\ D_{feature} &= [\sum W_{f\_1} \circ h_1, \sum W_{f\_2} \circ h_2, \sum W_{f\_N} \circ h_N] \end{aligned} \tag{7}$$

where $\circ$ is the hadamard multiplication, that means the weight is applied to the each feature element of h_i. To some extent, it reflects the contribution of different features to emotion classification. The general method combines the outputs of the last time step of each LSTM layer as the input of dense layer. The time method means the attention on the time dimension while the feature means that on the feature dimension.

3. Experiments and Analysis

3.1 Experimental Environment

The experiments are carried on the eNTERFACE [11] and IEMOCAP [12] corpus. The eNTERFACE is an audio and video emotion corpus in English, recorded from 43 speakers from 14 countries, and classifies samples based on the following 6 emotions: anger, disgust, fear, happy, sad and surprise. Only the audio data from this corpus is used in this

Table 1 Parameters of model

Parameters	Values/enterface	Values/IEMOCAP
Eta	1e-3, beta2 = 0.7	1e-4, beta2 = 0.7
Batch size	128	64
Epochs	1500	30000 steps
Lstm cells	[512, 256]	[256, 256]
Dense layers	[256/256*2, 128]	[256/256*2, 128]
Softmax layers	[128, 6]	[128, 5]
L2	1e-4	1e-4

work. 1260 valid speech samples are obtained for the emotion recognition study, of which 260 samples are used as the test set. IEMOCAP consists of about 12 hours of audiovisual data (speech, video, facial motion capture) from two recording scenarios: scripted play and improvised speech, in which 5 categories of emotion (anger, excited, frustrated, neutral and sad) with most samples (6355 samples in total) are selected as the research object.

In order to ensure the consistency of the comparison experiment, all super parameters adopted on the same corpus are exactly the same, as shown in Table 1. In dense layer, there are two kinds of parameters. [256, 128] represents the general LSTM network, while [256 ∗ 2, 128] is the LSTM network with skip-connections.

3.2 Features

The ComParE openSMILE features proposed by Schuller have been widely used in speech emotion recognition [13]. Based on these features, the frame-level speech features (i.e., the features without statistical functionals.) are directly used for emotion classification. The basic reasons are: (1) the fixed-length feature calculation of the statistical functional loses much information from the original speech, such as time information. (2) Hinton [14] believed that deep learning has the ability to automatically learn feature changes, and can learn deep features related to tasks from the underlying speech features. Thus, the frame level feature appears more suitable as an input to the deep learning network.

3.3 Convergence Property

In order to verify the convergence and stability of the proposed algorithm, the convergence curves of Enterface and IEMOCAP corpus are given in Fig. 2 and Fig. 3 respectively. The abscissa represents the number of steps of the iteration, and the ordinate is the value of unweighted average recall (UAR) that is used as a measure of performance in emotion recognition [13]. The initial slope of the curve reflects the convergence speed of the model, while the fluctuation degree of the curve reflects the stability of the model. As shown in figures, the feature attention model has the fastest convergence rate than the other models, especially on eNTERFACE corpus. In addition, the UAR of the proposed algorithm converges stably to 80% and 48% on eNTERFACE and IEMOCAP corpus respectively, while the traditional method converges stably to 70% and 43%, which means the proposed algorithm improves the performance of emotion recognition. Comparing with the fluctuation of the convergence curve without attention, the attention-weighted models not only improve the UAR and accelerate the convergence speed, but also do not sacrifice the stability of models.

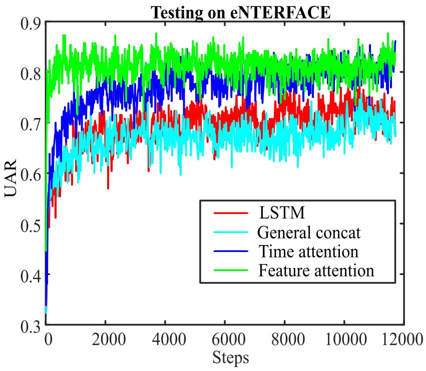

Fig. 2 Convergence curves on eNTERFACE

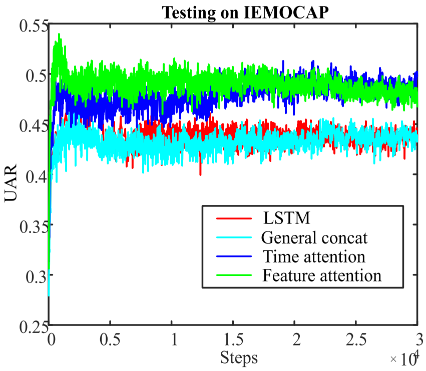

Fig. 3 Convergence curves on IEMOCAP

Table 2 Results of eNTERFACE

Model	Ang	Dis	Fea	Hap	Sad	Sur	UAR
LSTM	0.95	0.52	0.81	0.89	0.67	0.85	0.78
General	0.84	0.71	0.70	0.81	0.80	0.70	0.76
Time	0.91	0.88	0.81	1.00	0.91	0.70	0.86
Feature	0.88	0.81	0.85	0.97	0.87	0.89	0.88

Table 3 Results of IEMOCAP

Model	Ang	Exc	Fru	Neu	Sad	UAR
LSTM	0.38	0.40	0.47	0.49	0.56	0.46
General	0.50	0.25	0.46	0.46	0.66	0.47
Time	0.47	0.38	0.52	0.58	0.57	0.51
Feature	0.55	0.48	0.43	0.59	0.71	0.54

3.4 Results of Experiments

The recall of each emotion category and the UAR of each model are calculated in this study. Table 2 and Table 3 are the optimal recognition results on eNTERFACE and IEMOCAP corpus respectively. Compared with LSTM model, UAR of general model has not been clearly improved, and even decreased by 2% on Enterface corpus because skip connection brings a potential risk that the redundant information at the bottom obscures the useful information at the top. For this reason, it is necessary to add weight to the output of different layers so that the effective information can obtain a large weight and avoid the interference of redundant information from the bottom layer.

Compared with the models without attention, the time-dimension weighted model increased by 10% and 4% on Enterface and IEMOCAP corpus respectively, while the feature-dimension weighted model increased by 12% and 7% respectively. The experiments demonstrate that it can effectively improve the performance of the model to introduce the attention weight to the output of each layer, especially the effect of weighting on the feature dimension.

4. Conclusion

In this study, the attention mechanism is applied to LSTM networks with skip-connections for speech emotion recognition. Two methods are proposed for adding weight to skip connections from time dimension and feature dimension. Experiments on eNTERFACE and IEMOCAP corpus demonstrate that the proposed algorithm not only effectively improves the performance of emotion recognition, but also accelerates the convergence speed of the model under the premise of ensuring stability. Moreover, the difference between weights solves the potential risk caused by skip connection, that is, the problem of covering the effective information of the upper layer with the redundant information of the lower layer. Further research work includes the following points: first, the attention-weighted algorithm in time dimension is more meaningful for the continuous emotions recognition. Secondly, this attention-based LSTM is expected to be conducted in more applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61871213, 61673108, 61571106) and Six Talent Peaks Project in Jiangsu Province (2016-DZXX-023).

References

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

[2] M. Wollmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, et al., "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," Rubber Chemistry & Technology, vol.24, pp.638–639, 2008.

[3] G. Keren and B. Schuller, "Convolutional RNN: An enhanced model for extracting features from sequential data," in International Joint Conference on Neural Networks, pp.3412–3419, 2016.

[4] 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," vol.40, pp.842–850, 2015.

[5] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," IEEE Trans. Multimedia, vol.19, pp.1245–1256, 2017.

[6] Z. Lin, M. Feng, C.N.D. Santos, M. Yu, B. Xiang, B. Zhou, et al., "A structured self-attentive sentence embedding," arXiv preprint arXiv:1703.03130, 2017.

[7] M.-T. Luong, H. Pham, and C.D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," Computer Science, pp.1412–1421, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Computer Vision and Pattern Recognition, pp.770–778, 2016.

[9] Y. Liu, S. Piramanayagam, S.T. Monteiro, and E. Saber, "Dense Semantic Labeling of Very-High-Resolution Aerial Imagery and LiDAR with Fully-Convolutional Neural Networks and Higher-Order CRFs," in IEEE Conference on Computer Vision & Pattern Recognition Workshops, pp.1561–1570, 2017.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Computer Science, 2014.

[11] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface05 audio-visual emotion database," Proc. 22nd International Conference on Data Engineering Workshops, 2006, p.8, 2006.

[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 & Evaluation, vol.42, no.4, pp.335–359, 2008.

[13] J. Deng, X. Xu, Z. Zhang, S. Fruhholz, and B. Schuller, "Semi-Supervised Autoencoders for Speech Emotion Recognition," IEEE/ACM Transactions on Audio Speech & Language Processing, vol.26, no.1, pp.31–43, 2018.

[14] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in IEEE International Conference on Acoustics, Speech and Signal Processing, pp.5884–5887, 2011.