DILATED RESIDUAL NETWORK WITH MULTI-HEAD SELF-ATTENTION FOR SPEECH EMOTION RECOGNITION

Runnan Li [1,2], *Zhiyong Wu*[1,2,4], *Jia Jia*[2], *Sheng Zhao*[3], *Helen Meng*[4]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Tsinghua University
[2]Dept. of Computer Science and Technology, Tsinghua University
[3]Search Technology Center Asia (STCA), Microsoft
[4]Dept. of Systems Engineering and Engineering Management, The Chinese University of Hong Kong
lirn15@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn

ABSTRACT

Speech emotion recognition (SER) plays an important role in intelligent speech interaction. One vital challenge in SER is to extract emotion-relevant features from speech signals. In state-of-the-art SER techniques, deep learning methods, e.g, Convolutional Neural Networks (CNNs), are widely employed for feature learning and have achieved significant performance. However, in the CNN-oriented methods, two performance limitations have raised: 1) the loss of temporal structure of speech in the progressive resolution reduction; 2) the ignoring of relative dependencies between elements in suprasegmental feature sequence. In this paper, we proposed the combining use of Dilated Residual Network (DRN) and Multi-head Self-attention to alleviate the above limitations. By employing DRN, the network can retain high resolution of temporal structure in feature learning, with similar size of receptive field to CNN based approach. By employing Multi-head Self-attention, the network can model the inner dependencies between elements with different positions in the learned suprasegmental feature sequence, which enhances the importing of emotion-salient information. Experiments on emotional benchmarking dataset IEMOCAP have demonstrated the effectiveness of the proposed framework, with 11.7% to 18.6% relative improvement to state-of-the-art approaches.

Index Terms— dilated residual network, multi-head self-attention, speech emotion recognition

1. INTRODUCTION

In human speech interaction, paralinguistic characteristics like emotions, intonations and styles are employed to convey the underlying intent of messages. Recognizing and interpreting to these paralinguistic characteristics can help intelligent spoken interaction systems to understand the underlying user intention, and further improve the user experience. Therefore, automatic speech emotion recognition (SER) has become a research focus in human-computer interaction field.

In the development of SER systems, a vital challenge is to extract emotion-relevant features from speech [1]. Traditionally, the most popular approach is to apply a series of statistical aggregation functions (e.g., mean, max, variance, etc) on acoustic features (e.g., pitch, energy, etc) extracted from speech, to produce a long statistical feature vector for emotion classification [2]. The produced feature vector can roughly describe the temporal variations of speech signals, which are assumed to be highly related to the underlying emotion. With the development of deep learning technology, automatic feature learning algorithms are proposed to learn task-oriented features for SER, such as Deep Neutral Networks (DNNs), Convolutional Neutral Networks (CNNs), Recurrent Neutral Networks (RNNs), and their variants [3, 4, 5, 6]. Particularly, being developed from perception mechanism of the living creatures, CNNs has strong ability to filter out task-irrelevant information from input speech, proving clear patterns for the emotion classifier for emotion inferring [7]. Same as the successful employment in automatic speech recognition (ASR) [8] and speaker identification [9], the using of CNNs has also achieved significant improvement on SER task comparing to conventional approaches [4, 10, 11].

In state-of-the-art SER systems [4, 10, 11], CNNs are applied over windows of acoustic features with progressive resolution reduction, to produce higher-level features for the upper emotion classifier. However, while the progressively downsampling provides strong capability in local context modeling and emotion-related patterns detecting, the temporal structure of speech will also gradually lose in this process. As temporal evolution of speech is assumed to be highly related to the emotions [12], such loss can hamper the effectiveness of the SER system. In addition, in human emotion perception, attention-capturing vocalizations can produce greater activation to cortex, and significantly affect the perception of other vocalizations [13]. Capturing these relative dependencies between vocalizations can help the network to focus on emotion-salient parts of speech, and selectively import emotion-relevant information in feature learning, providing more discriminative representation for the emotion classifier. However, in CNN-oriented state-of-the-art methods, such dependencies are less considered.

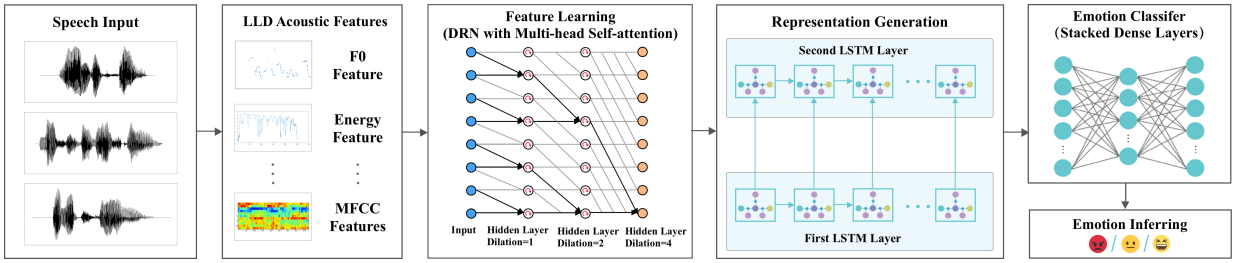

Fig. 1. The proposed framework uses LLDs as input, employs DRN with Multi-head Self-attention (represented by red recurrences) for feature learning, generates utterance-level representation with LSTM, and infers emotion with DNN based classifier.

In this paper, we propose the use of Dilated Residual Network (DRN) [14] and Multi-head Self-attention [15] for SER, to alleviate the loss of temporal structure and model the relative dependencies between suprasegmental features. Comparing to CNN based approaches, DRN inherits the properties of residual network, keeping the temporal structure of input signals through the network; and with dilation, the network can compensate the reducing of receptive field, proving strong ability in modeling local context. Self-attention is an attention mechanism relating different positions of input sequence; with multi-head mechanism, the function can further jointly attend information from different representation subspaces to improve the modeling performance. By applying Multi-head Self-attention in DRN, the network can model the underlying dependencies between different positions of suprasegmental features, proving selectively focusing on emotion-salient information in feature learning process.

The overall structure of the proposed SER framework is depicted in Fig.1. In inference, Low-Level Descriptors (LLDs) [2] are extracted from speech signal and used as the input. The DRN with Multi-head Self-attention is employed to generate suprasegmental features with high temporal resolution and selective attention. Long Short-Term Memory Recurrent Neural Network (LSTM) is employed to produce the utterance-level representation vector from the suprasegmental features sequence, and a DNN based classifier with stacked dense layers is employed to infer emotion with the generated representation.

2. METHODOLOGY

In this paper, we propose the use of DRN with Multi-head Self-attention for feature learning in SER, which can alleviate the loss of temporal structure and capture the relative dependencies of elements in the progressively feature learning.

2.1. Dilated Residual Network

The Dilated Residual Network (DRN) employed in this work is developed from [14], with starting point of residual network presented in [16], and consists with five groups of 1-D temporal convolutional layers. In residual network, for i^{th} layer in group $\mathcal{G}_i^l$, where $l = 1, ..., 5$, the output is calculated as

$$(\mathcal{G}_i^l * f_i^l)(\mathbf{p}) = \sum_{\mathbf{a}+\mathbf{b}=\mathbf{p}} \mathcal{G}_i^l(\mathbf{a})\, f_i^l(\mathbf{b}) \tag{1}$$

where f_i^l is the filter associated with $\mathcal{G}_i^l$, and the domain of $\mathbf{p}$ is the feature map in $\mathcal{G}_i^l$. A nonlinearity is in the following, which is omitted in the equation for clear declaration. Striding with a factor of 2 is employed in the first layer of each group for downsampling, reducing the temporal resolution and increasing the receptive field of convolutional layers.

To increase temporal resolution in higher layers for retaining temporal structure, a simple approach is to remove striding from some interior layers. However, this will hamper the network from exploiting contextual information while removing subsampling (striding) will cause the reduction of receptive field in subsequent layers. Since the context is important in human speech emotion perception, such loss will cause performance degradation in SER. To alleviate this problem, dilated convolution [17] is employed in DRN to compensate the reduction in receptive field from striding removing. For layers with k-dilated convolution, the output is calculated as

$$(\mathcal{G}_i^l *_k f_i^l)(\mathbf{p}) = \sum_{\mathbf{a}+k\mathbf{b}=\mathbf{p}} \mathcal{G}_i^l(\mathbf{a})\, f_i^l(\mathbf{b}) \tag{2}$$

The structure of the employed DRN is shown in Fig.2(b). In $\mathcal{G}_i^4$ and $\mathcal{G}_i^5$, the striding layers are eliminated and the convolutional layers are replaced with 2-dilated, 4-dilated convolutions, respectively. Thus, for input acoustic feature sequence with 10 msec frame length, the extracted suprasegmental features are with a granularity of 40 msec comparing to 160 msec in conventional CNN structure, which provides more detailed temporal variant information to representation production.

2.2. Multi-head Self-attention

Multi-head Self-attention mechanism [15] is proposed to model the relative dependencies between elements with different positions in sequence, enhancing the information importing from emotion-salient parts of speech. Developed from self-attention, Multi-head Self-attention maps the input sequence and a set of key-value pairs to a weighted output,

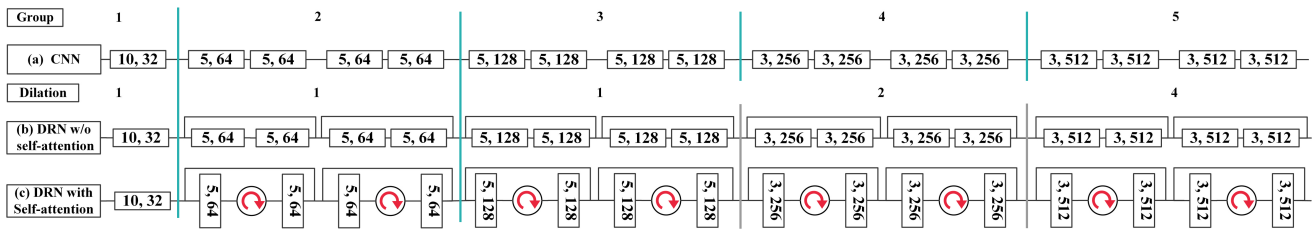

Fig. 2. Changing the CNN to DRN with Multi-head Self-attention. Group of layers are separated by the bold lines, and striding downsampling with a factor of 2 is processed after the blue lines. In DRN, 2-dilated and 4-dilated convolutional layers are employed in group 4 and 5 ($\mathcal{G}^4$ and $\mathcal{G}^5$), and the red recurrences represent the Multi-head Self-attention computing.

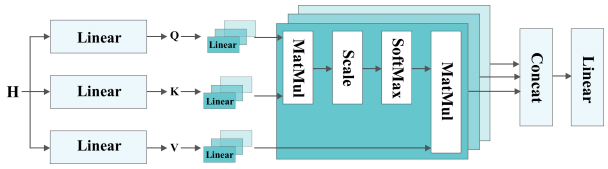

Fig. 3. The parallel structure of Multi-head Self-attention.

where the weights assigned are computed by a compatibility function using the input sequence and corresponding key.

As depicted in Fig.3, for frames in given hidden sequence H, the self-attention computes queries, keys, values of dimension d_k, d_k, d_v with linear projections. By packing them into matrices Q, K, V, the attention output is calculated as

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

To exploit the information from different representation subspaces at different positions, multi-head attention is further proposed to perform multiple attention function r times to generate queries, keys, values matrices Q_i, K_i, V_i from $i = 1, ..., r$. This mechanism has reported with higher effectiveness in producing attention representation in [15]. The Multi-head Self-attention is calculated as

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, ..., head_r)W^O \\ where\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{4}$$

where W_i^Q W_i^K, W_i^V are the weight matrices in parallel attentions with dimension d_k/r, d_k/r , d_v/r, respectively. W^O is the output weight matrix with dimension d_o.

In this work, as depicted in Fig.2(c), Multi-head Self-attention is applied in residual blocks in DRN, of each with $r = 4$ parallel heads. With the construction, the network can exploit the relative dependencies of elements at each group level, progressively enhancing the emotion-relevant information importing in feature learning procedure.

2.3. Speech Emotion Recognition

LSTM is employed to produce the utternance-level representation. For the suprasegmental feature sequence $M = (m_1, ..., m_n, ..., m_N)$, the network computes hidden sequence $h = (h_1, ..., h_n, ..., h_N)$ from $n = 1$ to N as

$$f_n = \sigma(W_f m_n + U_f h_{n-1} + b_f) \tag{5}$$
$$i_n = \sigma(W_i m_n + U_i h_{n-1} + b_i) \tag{6}$$
$$o_n = \sigma(W_o m_n + U_o h_{n-1} + b_o) \tag{7}$$
$$c_n = f_n \circ c_{n-1} + i_n \circ tanh(W_c m_n + U_c h_{n-1} + b_c) \tag{8}$$
$$h_n = o_n \circ tanh(c_n) \tag{9}$$

where σ is the Sigmoid activation function, f, i, o and c are the *input gate*, *forget gate*, *output gate* and *memory cell* activation vectors respectively, W, U and b items are the weight matrices and bias vectors of each gate. The last hidden output h_N is used as the utterance-level representation, and fed into the classifier to infer emotion.

3. EXPERIMENTS

3.1. Experimental Setup

Database. IEMOCAP [18] database is employed for performance assessment, of which contains 12 hours English conversations, segmented and categorized into utterances with 9 emotion classes. Same to the reported procedure in state-of-the-art techniques, utterances in 'exciting' class are combined to the happy class in evaluation, to form a four-class database labeled with $\{happy, angry, sad, neutral\}$, each class contains $\{1636, 1103, 1084, 1708\}$ utterances respectively.

Features. As suggested in the series of computational paralinguistic challenges (ComParE) [2], the TUMs open-source openSMILE [19] feature extractor is employed to extract LLDs from utterances, which contains 12-dimensional Mel-frequency cepstral coefficients (MFCCs), 1-dimensional logarithmic energy, voicing probability, harmonic-to-noise ratio (HNR), logarithmic fundamental frequency (LF0) and zero-crossing rate. Resulting in 17-dimmensional features are extracted for input frames, of which has 40 msec frame window length and 10 msec frame intervals.

Implementation and training. TensorFlow [20] is employed to implement the proposed framework as well as the comparisons. Cross-entropy metrics is employed as the

Table 1. Comparison results on IEMOCAP. Unweighted accuracy (UA), and F1-measure score (F1) are the higher the better. (*: the original performance reported in paper.)

	Parameters	Reported* UA	UA	F1
[Xia,2017]	9.5M	60.1%	60.3%	0.599
[Poria,2016]	9.3M	61.3%	60.5%	0.602
[Poria,2017]	9.4M	57.1%	56.8%	0.571
[Mirsamadi,2017]	9.6M	58.8%	59.7%	0.589
The proposed	9.9M	-	**67.4%**	**0.671**

loss function, Adam [21] optimization algorithm with an initial learning rate at 10^{-4} is employed in training. Back-propagation through time (BPTT) is employed to train the LSTM oriented models.

Metrics. To assess the performance of the implemented systems, unweighted accuracy (UA) and F1-measure [22] are employed. The UA is defined as the mean of accuracies for different emotion categories, and all the reported experimental results were based on 10-fold cross validation.

3.2. Comparison to State-of-the-art Approaches

Four representative approaches with reported performance on IEMOCAP are selected as comparisons. In implementation, to ensure the parameters consistency, the numbers of filters/units of convolution/dense layers in comparison methods are balanced.

1) [Xia, 2017] proposed the use of a series of statistics functions on acoustic features for representation production, and DNN based classifier for SER [3].
2) [Poria, 2016] proposed the use of CNN for representation learning, and multiple kernel learning classifier to infer speech emotion [4].
3) [Poria, 2017] employs conventional LSTM model to capture inner contextual information in speech for emotion recognition [5].
4) [Mirsamadi, 2017] employs local-attention enhanced RNN for feature learning and emotion inferring [6].

Experimental result. As shown in Table.1, our implementations of state-of-the-art approaches have achieved close performance to the original reported results. With the combining usage of DRN and Multi-head Self-attention in feature learning, the proposed framework has achieved significant relative improvement comparing to state-of-the-art approaches, +11.7% to [Xia, 2017], +11.4% to [Poria, 2016], +18.6% to [Poria, 2017], +12.9% to [Mirsamadi, 2017].

3.3. Component Contribution Research

To further discuss the contribution of individual components, four comparison systems with different combining of components or measures are implemented:

Table 2. Experimental results for component contribution research. Unweighted accuracy (UA), and F1-measure score (F1) are the higher the better.

	Residual	Dilated	Self-attention	UA	F1
Baseline	No	No	No	60.3%	0.599
S1	Yes	No	No	61.8%	0.621
S2	Yes	Yes	No	63.1%	0.637
S3	Yes	Yes	Single-head	66.9%	0.670
S4	Yes	Yes	Multi-head	**67.4%**	**0.671**

1) Baseline system, conventional CNN based approach[4].
2) System 1 (S1), residual network structure is employed to alleviate the loss of temporal structure.
3) System 2 (S2), dilation is employed to increase the receptive field of the residual convolution layers.
4) System 3 (S3), conventional self-attention is employed for relative dependencies modeling in feature learning.
5) System 4 (S4), multi-head attention mechanism is employed to enhance the system performance.

Experimental result. As shown in Table.2, started with the baseline system, when applying the residual network structure, the S1 has gained +2.5% relative improvement, when applying the dilated convolutional layers, the S2 has gained +2.1% relative improvement, when applying single-head self-attention, the S3 has gained +6.0% relative improvement, when enhancing with multi-head mechanism, the proposed framework has gained +0.7% relative improvement.

4. CONCLUSION

In this paper, we proposed the combining use of Dilated Residual Network and Multi-head Self-attention for feature learning in speech emotion recognition framework. With the residual structure, the network can maintain the temporal structure of input acoustic features in the regressive convolution processes. By employing dilated convolutional layers, the network can compensate the receptive field reduction caused by removing striding layers. Using Multi-head Self-attention, the network can model the relative dependencies between different positions in the suprasegmental feature sequence, denoting selectively focusing on emotion-salient parts of speech in feature learning. Experiments demonstrate the contribution and effectiveness of employing these techniques in SER, with significant improvement comparing to state-of-the-art approaches.

Acknowledgment. This work was conducted when the first author was an intern at Microsoft, and is supported by joint research fund of National Natural Science Foundation of China - Research Grant Council of Hong Kong (NSFC-RGC) (61531166002, N CUHK404/15), National Natural Science Foundation of China (61433018, 61375027) and National Social Science Foundation of China (13&ZD189).

5. REFERENCES

[1] Marie Tahon and Laurence Devillers, "Towards a small set of robust acoustic features for emotion recognition: challenges," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 1, pp. 16–28, 2016.

[2] Björn Schuller, Stefan Steidl, Anton Batliner, Elika Bergelson, Jarek Krajewski, Christoph Janott, Andrei Amatuni, Marisa Casillas, Amdanda Seidl, Melanie Soderstrom, et al., "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 2017, pp. 3442–3446.

[3] Rui Xia and Yang Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on Affective Computing*, , no. 1, pp. 3–14, 2017.

[4] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 2016, pp. 439–448.

[5] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, vol. 1, pp. 873–883.

[6] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2227–2231.

[7] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al., "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[8] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[9] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[10] Wootaek Lim, Daeyoung Jang, and Taejin Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Signal and information processing association annual summit and conference (APSIPA), 2016 Asia-Pacific*. IEEE, 2016, pp. 1–4.

[11] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5200–5204.

[12] Sylvie JL Mozziconacci and Dik J Hermes, "Expression of emotion and attitude through temporal speech variations," in *Sixth International Conference on Spoken Language Processing*, 2000.

[13] Annett Schirmer and Ralph Adolphs, "Emotion perception from face, voice, and touch: comparisons and convergence," *Trends in cognitive sciences*, vol. 21, no. 3, pp. 216–228, 2017.

[14] Fisher Yu, Vladlen Koltun, and Thomas A Funkhouser, "Dilated residual networks.," in *CVPR*, 2017, vol. 2, p. 3.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[18] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[19] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[20] Martín Abadi, Ashish Agarwal, and et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[21] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] David Martin Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.