Monaural Speech Separation Joint With Speaker Recognition: A Modeling Approach Based On The Fusion Of Local And Global Information

Shuang-qing Qian, Jing-jing Chen, He-ping Song, Qi-rong Mao
School of Computer Science and Communication Engineering, Jiangsu University

一等奖

Abstract

In recent years, the dominant speech separation methods model the sequence gradually to encompass global information with the time steps accumulated. However, due to the limited memory capacity of the model, the sequence information in the later position occupies a large proportion, making it impossible to form good interactions between sequences that are farther away. In this paper, we propose a modeling approach Dual-Path Recurrent Neural Network with Transformer (DPRNN-Transformer), which is based on the fusion of local and global information. Through this method, the interrelationships between sequences can be directly established. And the proposed model can effectively merge local and global information in the high-dimensional space by combining both sequence masking and non-masking, thus solving problems such as sequence information forgetting in speech separation area. In addition, in order for the model to extract as much speaker-related feature information as possible, we add the auxiliary task of speaker recognition after speech separation. By co-training with speaker recognition, the speech separation module will be constrained by the additional Triplet loss, thus incorporating speaker information to facilitate separation. Experimental results on WSJ0-2mix dataset show that our proposed method greatly improves the performance of speech separation.

Proposed DPRNN-Transformer

1. Framework

Depicted in Fig.1, our proposed model consists of four stages: encoder, block processing, decoder and speaker recognition. First, the encoder transforms the mixed waveform into their corresponding representations in an intermediate feature space, similar to the STFT in time-frequency separation. And the block processing handle the transformed representations to estimate the multiplicative function (mask) for each source at each time step. Then, the source waveform is reconstructed by transforming the masked encoder feature using a decoder module. Finally, we feed the evaluated speech into the speaker recognition module to extract speaker embedding.

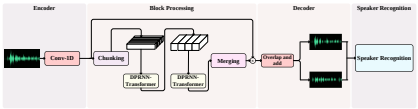

Fig.1 Framework of speech separation with DPRNN-Transformer

2. The Architecture Of DPRNN-Transformer

The original transformer structure is shown in Fig.2. It consists mainly of encoders and decoders, which are composed of multi-head attention and feedforward neural network. The encoder maps an input sequence of symbol representations to a sequence of continuous representations. And the decoder then generates an output sequence of symbols one element at a time based on the output of the encoder In this structure, multi-head attention is applied to a sequence of state vectors and transforms each state into a weighted average over all the states in the sequence, with more relevant states being given more influence.

Since Transformer consists entirely of attention, it is unable to learn the position relations of the sequence, and thus position encoding is added to compensate for this. Meanwhile, RNNs are effective at keeping track of positional information. Therefore, we consider stacking multiple Bi-LSTMs instead of the original position encoding to capture the sequence's position information and interrelationships. Then the Transformer encoder captures the global information, and the local information is modeled by the masked decoder step by step. Finally, global and local information are fused in a high-dimensional feature space. Our proposed DPRNN-Transformer architecture is shown in Fig.3

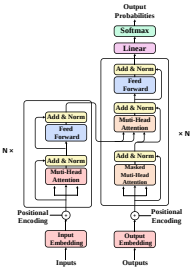

Fig.2 Architecture of original Transformer

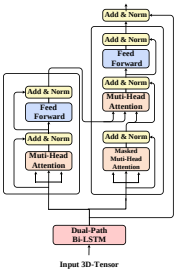

Fig.3 Architecture of proposed DPRNN-Transformer

In the architecture of DPRNN-Transformer, the output of DPRNN **u** is fed into the transformer. It should be noted that the original transformer is proposed for machine translation tasks, and thus the inputs to the encoder and decoder are the statements to be translated and the translated statements, respectively. In contrast, the inputs to the transformer encoder and decoder are the same in our model. We feed **u** into the Transformer encoder's multi-head attention, thus associating different positions of input sequences to calculate latent representations. The multi-head attention consists of **h** parallel self-attention layers, each of which is called a head. For each head, query, key and value should be mapped with linear layers before attention calculation. The output of these **h** attention heads will be concatenated and then input into the last linear layer for integration. It can be formulated as follows:

$$\mathbf{o}^i = \mathbf{softmax}\left(\frac{(\mathbf{W}_q^{(i)}\mathbf{s})\cdot(\mathbf{W}_k^{(i)}\mathbf{s})^T}{\sqrt{\mathbf{d}}}\right)\mathbf{W}_v^{(i)}\mathbf{s},\ \mathbf{i=1,\ \ldots,\ h}$$

$$\mathbf{o} = \mathbf{W}_O[\mathbf{o}^1,\ldots,\mathbf{o}^h]$$

The masked Transformer decoder is different from the encoder in that it masks the input, so that the attention mechanism can only focus on the sequence before the current time step rather than the whole sequence. The formula is slightly different from the above as follows:

$$\mathbf{o}^i = \mathbf{softmax}\left(\frac{(\mathbf{W}_q^{(i)}\mathbf{s})\cdot(\mathbf{W}_k^{(i)}\mathbf{s})^T}{\sqrt{\mathbf{d}}}+\mathbf{M}\right)\mathbf{W}_v^{(i)}\mathbf{s},\ \mathbf{i=1,\ \ldots,\ h}$$

$$\mathbf{M} = \begin{bmatrix} 0 & -\infty & \cdots & -\infty \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

The Transformer decoder then sends the output of the masked sequence together with the encoder's output to another multi-head attention, resulting in the interaction of local and global information of the sequence at different time steps. Through the interaction of encoder and decoder, the model is able to make optimal decisions based on the accumulated local information and global information in the current time step. Thus, the local and global information is further fused in the high-dimensional space, improving the model's ability to fit sequences.

Experiments

1. Dataset

We evaluate our proposed model on two-speaker speech separation problem using WSJ0-2mix dataset, which contains 30 hours of training and 10 hours of validation data. The input mixtures are generated by randomly selecting utterances of different speakers from WSJ0 training set, and mixing them at random signal-to-noise ratios (SNR) between -5 dB and 5 dB . 5 hours of evaluation set is generated in the same way, using utterances from 16 unseen speakers in WSJ0 validation set and evaluation set. It should be noted that WSJ0-2mix does not include speaker labels. Hence, we add the corresponding speaker labels into WSJ0-2mix for joint training.

2. Configurations

The networks are trained for 100 epochs on 4-second long segments. The initial learning rate is set to 1e-3 and decays by 0.98 for every two epochs. Early stopping is applied if the loss of the validation set no longer drops for 10 consecutive epochs. Adam is used as the optimizer. Due to the limitation of GPU memory, we stack 1 Transformer structure with h=4 parallel attention layers after DPRNN.

3. Results

Table I lists the average scores of the well-known speech separation methods in recent years on SI-SNR and SDRi evaluation standards, and compares them with our proposed methods. Obviously, our result is superior to the state-of-the-art approach DPRNN on WSJ0-2mix corpus, and this further proves the effectiveness of our proposed modeling method based on local and global information fusion for the information forgetting problem, caused by time-step progressive modeling in the field of speech separation.

Table I Comparison with other methods on WSJ0-2mix

Method	SI-SNR(dB)	SDRi(dB)
DPCL++ [25]	10.8	-
uPIT-BLSTM-ST [9]	-	10.0
Deep Attractor [10]	10.5	-
ADANet [26]	10.4	10.8
WA-MISI-5 [27]	12.6	13.1
Conv-TasNet-gLN [13]	15.3	15.6
Conv-TasNet + MBT [28]	15.5	15.9
Deep CASA [29]	17.7	18.0
FurcaNeXt [30]	-	18.4
DPRNN [18]	18.8	19.0
DPRNN-Transformer	**19.5**	**19.7**

Conclusion

In this paper, we investigate the effectiveness of Transformer, and successfully introduce it into DPRNN to realize the fusion of local and global information for end to end monaural speech separation. The experimental results show that our proposed method effectively solves the problems of Transformer's inability to handle very long sequences and greatly improves the separation performance on the public WSJ0-2mix data corpus. In addition, by adding the assistant task of speaker recognition, the feature extracted by the separation module is equipped with speaker information, which further improves the performance of separation.

References

[1] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 46–50.

[2] Ziqiang Shi, Huibin Lin, Liu Liu, Rujie Liu, Shoji Hayakawa, Shouji Harada, and Jiqing Han, "End-to-end monaural speech separation with multi-scale dynamic weighted gated dilated convolutional pyramid network," in Interspeech, 2019, pp. 4614–4618.

Contact

Shuang-qing Qian
0086-17805015817
1017314232@qq.com