MULTIPLE CRNN FOR SELD

Technical Report

Congzhou Tian

Peking University
Wangxuan Institute of Computer Technology
No.128, Zhongguancun North Street, Haidian District, Beijing, P,R,China
tcz@pku.edu.cn

ABSTRACT

In this task, we use multiple CRNN for SELD. Firstly, there is a CRNN to predict the number of sound events at the same time. A SED CRNN is used to predict the current sound events given the activated number result. After that, we train a DOA1 CRNN specifically for frames with single active event and a total DOA CRNN for frames with more active events. We think training with separate network is helpful for both SED and DOA tasks and our results are proved better than the baseline method on the development dataset.

Index Terms— CRNN, SELD

1. INTRODUCTION

Sound event detection(SED) is a task to detect the onset and offset times for each sound event in an audio recording and associate a textual descriptor, i.e., a label for each of these events[1]. In the most simple case, there is at most one sound event at the same time and detection is done by analyzing and matching sounds' characteristics. But in real life recordings, overlapping sound events make it harder to detect each or desired sound event. Convolutional Recurrent Neural Networks(CRNN), which integrates the strengths of both CNN's and RNN's, has been proposed and proved to be better than both CNN and RNN models on the SED problem in [1].

In [2], authors propose to use a single CRNN to do both SED and DOA(Direction of Arrival) jobs and prove it's successful. In [3], a consistent way of measuring the joint performance of the SELD system is proposed, which are the components of evaluation metrics of DCASE2020 task3.

2. METHODOLOGY

We believe that SED and DOA are two different tasks. Judging by our common sense, DOA task with microphones array should be easier to handle than SED. Baseline CRNN method can work well only when the DOA output comes from the right sound event channel, which means the DOA output actually has mastered the SED function. In order to let the DOA function be optimized(which is helpful for the F_{20} criterion) alone, we try to use CRNN to do DOA work with the SED influence as less as possible.

We firstly train a NOA(number of arrivals) network using CRNN to obtain the number of current activated sound events. Another CRNN is used to predict the current sound events given the activated number result. The first one or two possible sound event are chosed depending on the NOA number we have. We find that the F1 score of SED is a bit worse than one-step SED. But it's worthy doing the two step works because you would obtain more reliable NOA results for DOA.

We train a DOA network(DOA1) for frames with single activated sound event and this DOA result is proved to be better than original baseline DOA result. Because the time resolution is quite high and limited working time, we didn't explore better DOA methods for two activated sound events. And the problem needs to be solved is to do alignment between the DOA and SED results. We choose to use the baseline network(DOA-ALL) to handle the two activated sound events methods. The illustration of system is shown in Fig. 1.

The detailed information of multiple CRNN we use is shown in Tab 1.

Table 1: Detailed Information of Multiple CRNN

infor	NOA	SED	DOA1	DOA-ALL
parameters	488211	508257	490326	513288
last layer activation	softmax	softmax	tanh	tanh
last layer units	3	14	3	42

3. EXPERIMENTAL RESULTS

We report our experimental results on the development dataset[4] in this section as shown in Tab 2. The testing data is ambisonic. Both the F_{20} and LR_{CD} metrics are improved. It shows that doing SELD work separately with multiple CRNN is better for results.

Table 2: Experimental Results on the Development dataset

Method	ER_{20}	F_{20}	LE_{CD}	LR_{CD}
baseline	0.72	37.4%	22.8	60.7%
Ours	0.72	40.6%	25.9	64.0%

We also tried one-step SED network and compare its results with SED with NOA results, as shown in Tab 3. Results show if the SED results are considered only, the one-step SED CRNN perform better than SED with DOA. We choose SED with DOA in our methods because the is number of sound events is needed.

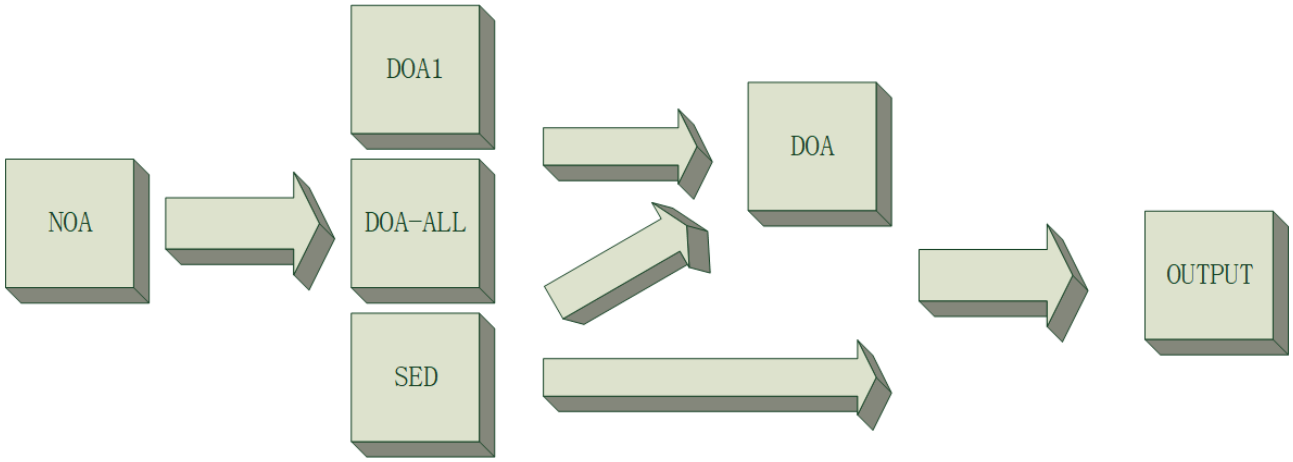

Figure 1: The entire system of our proposed method

Table 3: one-step SED compared with SED with NOA results

Method	LR_{CD}
one-step	64.0%
SED with NOA	67.6%

4. CONCLUSION

CRNN network is an appropriate architecture for SELD task. However, we think it's better for both tasks if they are trained separately. We train multiple CRNN for SED and DOA tasks. Comparing with the baseline method, our results are better on F_{20} and LR_{CD} metrics on the development dataset.

5. REFERENCES

[1] E. akr, G. Parascandolo, T. Heittola, H. Huttunen and T. Virtanen, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1291-1303, June 2017. doi: 10.1109/TASLP.2017.2690575

[2] Adavanne, Sharath , et al. "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks." IEEE Journal of Selected Topics in Signal Processing (2018):1-1.

[3] Mesaros, Annamaria and Adavanne, Sharath and Politis, Archontis and Heittola, Toni and Virtanen, Tuomas, "Joint Measurement of Localization and Detection of Sound Events", in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), Oct 2019.

[4] Politis, Archontis and Adavanne, Sharath and Virtanen, Tuomas, "A Dataset of Reverberant Spatial Sound Scenes with Moving Sources for Sound Event Localization and Detection", https://arxiv.org/abs/2006.01919