DCASE 2021 TASK 3: SELD SYSTEM BASED ON RESNET AND RANDOM SEGMENT AUGMENTATION

Technical Report

Zijun Pu

Faculty of Imformation Engineering and Automation,
Kunming University of Science and Technology,
Kunming, China
ppzzj123@163.com

Jisheng Bai, Jianfeng Chen

School of Marine Science and Technology,
Northwestern Polytechnical University,
Xi'an, China
baijs@mail.nwpu.edu.cn,
cjf@nwpu.edu.cn

ABSTRACT

This technical report describes our system proposed for Sound Event Localization & Detection (SELD) task of DCASE 2021 challenge [1]. In our approach, Resnet architectures are used in the task as main network for SELD, and GRU is used after the Resnet for catching temporal relationship of acoustic features. Moreover, a data augmentation method called random segment augmentation is adopted during training. Firstly, the original sound recordings which only contain single event sound are cut into 100ms length pieces. Secondly, all the sound pieces are shuffled and randomly combined for generating new recordings. Finally, our proposed system is evaluated on the development dataset of task3 and it achieve better performance than baseline.

Index Terms— Sound Event Localization & Detection, random segmentation, data augmentation

1. INTRODUCTION

The complex sound environment puts forward higher requirements on the robustness of the environmental sound recognition and positioning system, and the SELD task is to design a stable and efficient algorithm system for this complex environment. The SELD task of DCASE 2021 is a developed version of the task in DCASE2020 [2].

In this report, we use a data augmentaion method based on mission requirements and propose a SELD system for detection and localization. In order to generate more data, it is necessary to slice and remix the original data to obtain the amplified dataset. In this way, the dataset can be increased many times. After that, it is necessary to extract the features of the sound data. This report uses the current mainstream log-Mel spectrogram as acoustic features, and the GCC-PHAT as the spatial feature input to network. In addition to the aforementioned data enhancement algorithms, specAugment [3] is also an effective data enhancement method. In addition, there are EDMA algorithms designed based on the aforementioned data enhancement algorithm, etc.

From another point of view, sound recognition is an image recognition process. Therefore, many network architectures used for machine vision can be used for reference, and sound is a signal with temporal context, so many networks used for natural language processing can also be used for reference. Similarly, this report uses the Resnet network borrowed from machine vision to design the network structure for SELD tasks and verify its performance.

The rest of the report is organized as follows. In Section 2, the proposed method is described in detail, including data augmentation, network frame. Evaluation results on development dataset is shown in Section 3. Conclusions are summarized in Section 4.

2. PROPOSED METHOD

In this section, the data augmentation algorithm and DNN model are mainly introduced. First, the development dataset is preprocessed, and then the processed data is used to train the DNN and the final result is output by the network. Therefore, we chose the data enhancement algorithm of slice reorganization [4].

2.1. Data augmentation

The data provided by the development dataset is not sufficient for training the deep network, so the data set needs to be augmented to obtain a larger data set. Thus, we chose the data enhancement method of slice reorganization. This data enhancement is to cut out the sound fragments of a single sound event from the original data set, and randomly mix them into the original data set to obtain the amplified data set. The advantage of this data enhancement method is that it can retain most complete audio clips where multiple sound events occur at the same time. This also means that the timing relationship of these sound clips will not be lost due to data enhancement. The data enhancement method using slice reorganization can theoretically amplify the original data many times. The following Figure 1 shows the processing diagram of the data enhancement algorithm.

2.2. Network

As the state-of-the-art model in the field of computer vision, Resnet [5] has received extensive attention in recent years. The residual block introduced can make the deep network have a better convergence effect. In view of its excellent performance in the image field, we introduce it to the task of Sound Event localization and Detection.

In this technical report, we propose a network framework based on Resnet-34. The following Figure 2 shows the network structure diagram we proposed. For the characteristics of the input network, we choose Log-mel spectrum and GCC-PHAT.

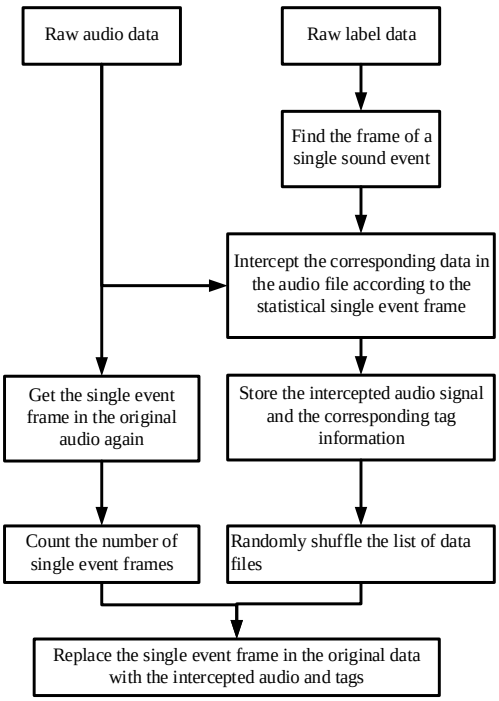

Figure 1: Processing diagram of proposed method.

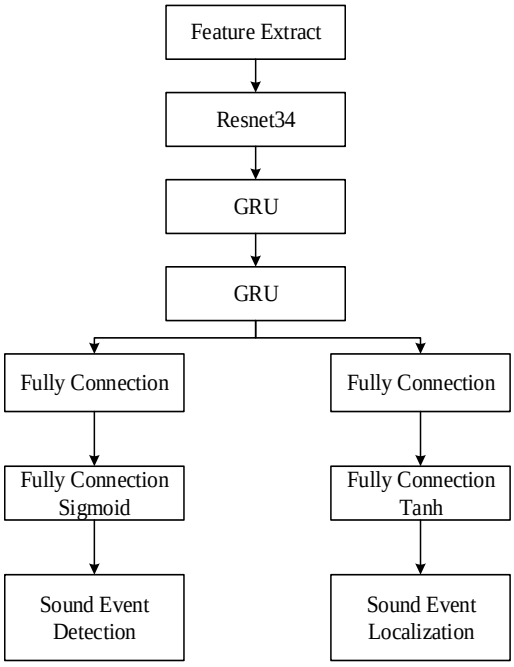

Figure 2: Proposed network structure.

This network is improved on the basis of the DCASE2021 baseline model [6, 7]. As the data of the DCASE2021task3 competition is more difficult to identify and locate than in previous years, the number of network layers used by the baseline model can no longer meet the demand in terms of capturing the details of audio features. Therefore, this technical report considers the introduction of Resnet-34 with a deeper network layer as a capture network for the details of the feature spectrum. At the same time, compared to last year, the sound data events of this year's competition are more natural. In order to ensure the efficiency of network training, the GRU part of the baseline model remains unchanged. The output part of the network is the same as that given by the baseline, and the sigmoid activation function and BCE loss function are used for the recognition part. For the positioning part, use Tanh activation function and MSE loss function.

3. RESULTS

In order to verify the effectiveness of the algorithm, we tested several systems including the limit baseline model on the development data set. The test results are shown in the following Table 1.

As can be seen from the above Table 1, in the baseline model, as the data enhancement factor increases, the system output result will not be improved. Data enhancement will reduce the effectiveness of the baseline model. The network structure designed based on Resnet also has problems similar to the baseline model, but when the data is doubled, the overall output of the model is the best. It can be seen that data enhancement can only improve the performance of the model to a certain extent. But overall, the network structure plays a decisive role in the performance of the entire system.

4. CONCLUSION

The proposed SELD system based on the Resnet-34 structure performs the best when the data is doubled increased. At the same time, the data enhancement algorithm of slice reorganization that we use has played a role in improving system performance to a certain extent. At the same time, according to the result on test split of the development dataset, we found that the effectiveness of this data enhancement algorithm changes according to the network structure. Data enhancement does not perform better on the baseline model, but the performance of our proposed method has been significantly improved.

5. REFERENCES

[1] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," *arXiv preprint arXiv:2106.06999*, 2021. [Online]. Available: https://arxiv.org/abs/2106.06999

[2] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020. [Online]. Available: https://https://arxiv.org/abs/2009.02792

[3] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[4] Q. Wang, H. Wu, Z. Jing, F. Ma, Y. Fang, Y. Wang, T. Chen, J. Pan, J. Du, and C.-H. Lee, "The ustc-iflytek system for

	ER_{20°	$F_{20^\circ}(\%)$	LE_{CD}	$LR_{CD}(\%)$
Baseline	0.73	24.6	32	40.6
Baseline Double-aug	0.70	24.5	35.3	40.2
Baseline Third-aug	0.75	20.5	38.2	31.8
Resnet	0.74	23.4	35.1	34.6
Resnet Double-aug	**0.67**	**32.6**	**30.5**	**46.5**
Resnet Third-aug	0.69	26.5	29.9	39.3

Table 1: Results on test split.

sound event localization and detection of dcase2020 challenge," DCASE2020 Challenge, Tech. Rep., July 2020.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, March 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8567942

[7] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, June 2021.