SOUND EVENT LOCALIZATION AND DETECTION BASED ON CRNN USING TIME-FREQUENCY ATTENTION AND CRISS-CROSS ATTENTION

Technical Report

Yin Xie[1,2], *Ying Hu*[1,2], *Yunlong Li*[1,2], *Shijing Hou*[1,2], *Xiujuan Zhu*[1,2], *Zihao Chen*[1,2], *Liusong Wang*[1,2], *Mengzhen Ma*[1,2],

[1] Xinjiang University, School of Information Science and Engineering, Urumqi, China
[2] Key Laboratory of Signal Detection and Processing in Xinjiang, Urumqi, China
{liyunlong, huying}@stu.xju.edu.cn

ABSTRACT

This report describes our systems submitted to the DCASE2022 challenge task 3: sound event localization and detection (SELD).We design a CRNN network based on asymmetric convolution mechanism with Time-Frequency Attention module(TFA) and Criss-Cross Attention module(CCA) which achieves great performance to deal with SELD in complex real sound scenes. On TAU-NIGENS SpatialSound Events 2022 development dataset, our systems demonstrate a significant improvement over the baseline system. Only the firstorder Ambisonics (FOA) dataset was considered in this experiment.

Index Terms— DCASE2022, Sound source localization,Sound event detection, Time-Frequency Attention

1. INTRODUCTION

Sound event localization and detection(SELD) consists of two subtasks which are sound event detection and direction-of-arrival estimation and aim to recognize the sound class, as well as estimating the corresponding direction of arrival (DOA), onset, offset of the detected sound event[1].SELD has many applications in surveillance [2], autonomous driving [3], home assistant systems and security applications.

In DCASE2022, the baseline model is similar to the one used in DCASE2021, which is based on a convolutional recurrent neural network (CRNN) stemming from the original SELDNet architecture [1] proposed by Adavanne, but improved with the activity-coupled Cartesian direction of arrival output representation (ACCDOA)[4]. The multi-ACCDOA [5]format not only enables the model to solve the cases with overlaps from the same class,but also improves SELD performance and reduces the network size in the Meanwhile.

In this report, We also propose a CRNN framework based on SELD-Net architecture for jointly detecting, classifying, and localizing the acoustic target classes using the FOA data.We adopt asymmetric convolution mechanism[6] with Time-Frequency Attention module (TFA) and Criss-Cross Attention module(CCA) [7]to process feature extracting more insuffificiently.Instead of conventional symmetric convolution, the TFA structure is design to process more and richer spatial and time-frequency features and increase feature diversity by asymmetric convolution.Recent studies[8] show lower layers often need more local information, while higher layers desire more global information.Inspired by this, we adopt CCA module to capture the longer global information.We conduct experiments on development dataset to verify the effectiveness of our proposed method.

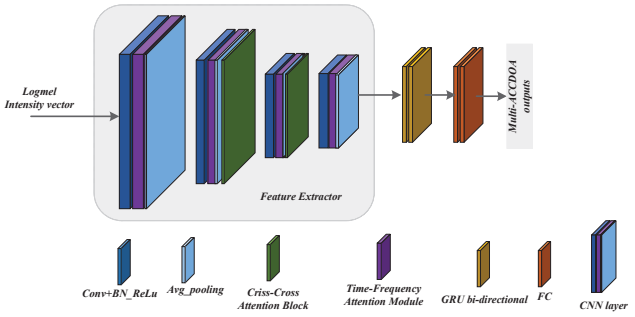

Figure 1: SELD network based on time-frequency attention

2. PROPOSED METHODS

2.1. Network architecture

In this paper, we apply CCA for the first time to SELD. The CCA can reduce the number of parameters and computational complexity while modeling global information.We design a network with asymmetric convolution mechanism[6] with time-frequency Attention module(TFA) and Criss-Cross Attention module(CCA) which achieve great performance to deal with SELD in complex real sound scenes.The network is capable of detecting and classifying the sound event classes active for each of the input frames along with their respective spatial location, and localizing the temporal activity and DOA trajectory up to two instances of each class simultaneously.

The overall architecture of our system is illustrated in Figure 1. The main difference compared with the baseline system is that the feature extractor module is substitute by TFA and CCA, which is proposed to model time-frequency , local and global dependencies of an audio sequence by combining asymmetric convolution neural networks,time-frequency Attention and Criss-Cross Attention[7]. The CCA module is depicted in Figure 2.The logmel spectrogram and sound intensity vector (SIV) extracted as the input feature maps are fed into the feature extractor firstly to extract high-level features. The feature extractor as depicted in Figure 2 and Figure 3.

After that, the time dimension is downsampled 8 times, and the frequency dimension is downbsampled 16 times. Then, Bidirectional Gated Recurrent Unit (Bi-GRU) is used to learn the temporal context information. This is followed by fully connected layers.

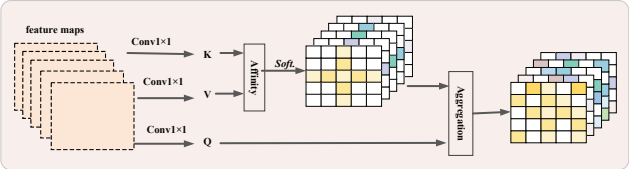

Figure 2: Criss-Cross Attention module

The CNN layers, which are also named as feature extractor layers in this paper, are constructed with 4 groups of one 2D CNN(Conv) and a Time-Frequency Attention module(TFA) layer with average-pooling after each of them. Each Convs' group consists of one 2D Conv, with a receptive field of 3 × 3, a stride of 1 × 1, and a padding size of 1 × 1,and a TFAM.The Convs' kernels filter across all of the channels of the input features and the feature maps from the last layer, hence are capable of learning interchannel information. CNN layers are able to learn local spatial information to better abstract the event-level information.Each 2D CNN is followed by a Batch Normalization layer[9] and a ReLU activation.In order to better capture the global time-frequency information and increase the correlation between the frame level, a Criss-Cross Attention module is used in seconde feature CNN layer and third CNN layer.After the feature extractor, the feature map has shape C ×T /8×F/16, where C is the number of output feature maps of the last CNN layer. It is then sent to a global average-pooling layer to reduce the dimension of F. After this, the feature map is reshaped to have shape T /8 × C and is fed to two bidirectional GRUs followed by a tanh activation and two fully-connected layers whichthe last one is followed by a tanh activtation.Finally,the feature map is sent to multi-ACCDOA model and predicted.

2.2. TFA module

First, two parallel 2D Conv with field of 3 × 1 and 1×3 respectively, to extract features along the time and frequency dimensions. After the features concate along the channel, the features are fused through a convolution with a kernel size of 1× 1 followed by a Batch Normalization layer and a ReLU activation.Then, the global average pooling layer is applied to each channel of the feature map, and the constant scalar of the feature channel can be obtained. Then C weight coefficients are obtained through two convolutions with a kernel size of 1× 1.Then, this set of weight coefficients and their complementary set of weight coefficients were multiplied by time feature and frequency feature respectively and then added. Finally, BN normalization and ReLu activation were used.

3. EXPERIMENTAL EVALUATION

3.1. Dataset

Development set of TAU-NIGENS Spatial Sound Events 2022 has two types of data, one is 4 channel directional microphone array (MIC) from tetrahedral array and the other one is first-order ambisonic (FOA) data.The dataset contains 1200 one-minute spatial mixtures are synthesized (synth-set) using SRIRs from 9 rooms

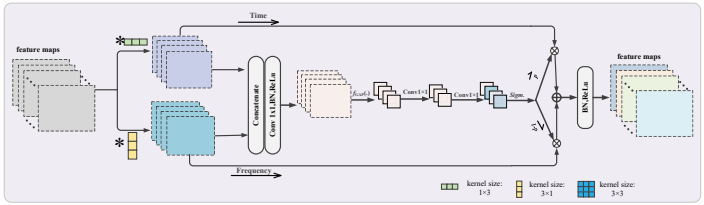

Figure 3: Time-Frequency Attention module(TFA)

in TAU and sound event samples sourced from FSD50K[10], 70 recording clips of 30sec-5min durations, with a total time of 2hrs, contributed by SONY (development dataset) and 51 recording clips of 1min-5min durations, with a total time of 3hrs, contributed by TAU (development dataset). In this paper, we only used the FOA (4-channels, 3-dimensional recordings) data format for the experiments. Contrary to the DCASE2021 Task3 dataset, the DCASE2022 dataset contains 13 different SED target classes, and up to five overlapping events may occur.

3.2. Experimental set

We choose the Adam optimizer with a learning rate of 0.0005, and the total training epochs are set to 100.Four-channel spectrograms for both formats are computed with 1024-point FFTs using a 40 msec hanning window and 20 msec hop length at 24kHz. Log-mel spectrograms are additionally extracted from the STFT ones at 64 mel-bands.All experiments were conducted on a GeForce RTX 2080 Ti .

3.3. Experimental results

As showing in table 1,our proposed model result outperform the DCASE 2021 baseline model.

model	$ER_{20^\circ} \downarrow$	$F_{20^\circ} \uparrow$(macro)	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$
baseline	0.71	21.0%	29.3	46.0%
our model	0.66	34.2%	22.9	57.7%

Table 1: Final results of the models submitted

4. REFERENCES

[1] 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, 2018.

[2] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE transactions on intelligent transportation systems*, vol. 17, no. 1, pp. 279–288, 2015.

[3] M. K. Nandwana and T. Hasan, "Towards smart-cars that can listen: Abnormal acoustic event detection on the road." in *INTERSPEECH*, 2016, pp. 2968–2971.

[4] 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 *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 915–919.

[5] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 316–320.

[6] X. Ding, Y. Guo, G. Ding, and J. Han, "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1911–1920.

[7] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.

[8] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 116–12 128, 2021.

[9] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[10] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.