ENSEMBLE OF ATTENTION BASED CRNN FOR SOUND EVENT DETECTION AND LOCALIZATION

Technical Report

Rong Xie[1], *Chuang Shi*[1], *Le Zhang*[1], *Yunxuan Liu*[2] *and Huiyong Li*[1]

[1]University of Electronic Science and Technology of China,
School of Information and Communication Engineering, Chengdu, China.
xierong@std.uestc.edu.cn, shichuang@uestc.edu.cn

[2]University of Electronic Science and Technology of China,
School of Electronic Science and Engineering, Chengdu, China.

ABSTRACT

This report describes submitted systems for sound event localization and detection (SELD) task of DCASE 2022, which are implemented as multi-task learning. Soft parameters sharing convolutional recurrent neural network (CRNN) with Split attention (SA), convolutional block attention module (CBAM) and coordinate attention (CA) are trained and ensembled to solve the SELD task. To generalize models, angle noise and mini-batch time-frequency noise are introduced, and mini-batch mixup, FOA rotation, frequency shift, random cutout and SpecAugment are adopted. Proposed systems have a better performance than the baseline system on the development dataset.

Index Terms— Sound event localization and detection, CRNN, attention mechanism, model ensemble, data augmentation

1. INTRODUCTION

The SELD task in DCASE2022 is desired to detect events' classes and their directions in the polyphony and reverberation scenarios. Comparing to DCASE 2021 Task 4, real recordings are added, which makes the events' time-frequency features and rebervation more variable. In order to capture the category and direction information of sound events in the noise and reverbreration enviroment, SA[1], CBAM[2], CA[3] are adopted in networks. To combat noise and interference, angle noise using Rodrigues' rotating changes the events' direction label in a contiguous event, and in a mini-batch, one's random time and frequency clips are added to another by multiplying a factor. Inspired by Yin Cao[4], CNN encoders with soft parameters sharing (soft-PS) are built. And the decoders are bidirectional GRU (Bi-GRU) and fully connected (FC) layers. Fig. 1. shows the network framework in this report.

2. ATTENTION BASED SELD SYSTEM

In this section, features, data augmentation, and attention (SA, CBAM and CA) based SELD CRNN models are described.

2.1. Features

In this report, only FOA recording format is used to train the models. Training clips are extracted to frame-wise log-mel spectrograms

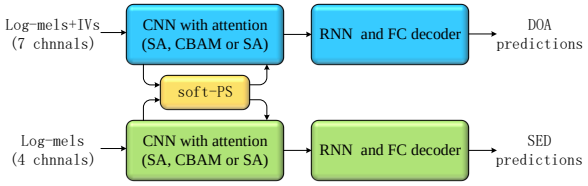

Figure 1: Soft parameters sharing network frameworks

($Log\text{-}mels$) and intensity vectors (IVs)[5]. STFT is computed firstly. Then, for the $Log-mels$,

$$Log\text{-}mels = mel\_w \bullet \begin{bmatrix} 20log(\|(stft\_spect_W\|) \\ 20log(\|(stft\_spect_Y\|) \\ 20log(\|(stft\_spect_Z\|) \\ 20log(\|(stft\_spect_X\|) \end{bmatrix}, \quad (1)$$

where mel_w is mel filter banks' weights, and W, Y, Z and X are channel indexes of FOA recording format, and $\bullet$ denotes dot product.

For the IVs, IV_{re} is firstly computed as

$$IV_{re} = \begin{bmatrix} Re(stft\_spect_W^* \circ stft\_spect_Y) \circ \|stft\_spect_W\| \\ Re(stft\_spect_W^* \circ stft\_spect_Z) \circ \|stft\_spect_W\| \\ Re(stft\_spect_W^* \circ stft\_spect_X) \circ \|stft\_spect_W\| \end{bmatrix}, \quad (2)$$

and $\circ$ denotes Hadamard product. To normalize the IV_{re}, the $norm\_term$ is computed as

$$norm\_term = \sqrt{IV_{re}(Y)^2 + IV_{re}(X)^2 + IV_{re}(Z)^2} \quad (3)$$

At last, IVs can be expressed as

$$IVs = mel\_w \bullet (IV_{re} \oslash norm\_term), \quad (4)$$

where $\oslash$ denotes element-wise division. For the SED branch of networks, only $Log\text{-}mels$ are used as input features. For the DOA branch of networs $Log\text{-}mels$ and IVs are concatenated to be the input features.

2.2. Data augmentation

Mini-batch mixup[9], angle noise, mini-batch time-frequency noise, FOA rotation[6], random cutout[7] and SpecAugment[8] are implemented to generalize the model. The indexes random permuted features add features with original indexes. If the same class exists for two segments, the addition operation will not be performed.

For angle noise, Rodrigues' rotating is applied in a contiguous event to change its DOA label,

$$\boldsymbol{v}_{rot} = \boldsymbol{v}cos\theta + (\boldsymbol{u} \otimes \boldsymbol{v})sin\theta + \boldsymbol{u}(\boldsymbol{u} \bullet \boldsymbol{v})(1 - cos\theta), \quad (5)$$

where $\boldsymbol{v}$ is the original DOA label represented by Cartesian coordinates, $\otimes$ denotes Kronecker product, $\boldsymbol{u}$ is a random unit vector, and θ is the angle to be rotated, which is uniformly distributed in $[-6°, 6°]$.

To bring extra reverberation and noise to the training data, time-frequency noise is applied in a mini-batch. The indexes of features are randomly permuted. Two time cutouts and tow frequency cutouts from indexes permuted features multiplied by a factor are added to original indexes features, and the factor is uniform distributed in $[0.05, 0.1]$.

For FOA rotation, $Swap(X, Z)$ and $Swap(Y, Z)$ are added compared to FOA rotation in [6].

2.3. Network architectures

Networks in this report adopt CRNN architecture in [5]. ResNet22 in PANNs[10] and soft-PS[4] are referenced to build the CNN encoder. There are two branches in the network, one for predicting SED, another for estimating DOA .

The network architecture is shown in Fig. 2. Input and output shapes of modules in the network are shown in square brackets. The shape of the network input features is [B, C, T, F]. C is the number of channel for input features, 4 for SED branch and 7 for DOA branch. B, T, F are batch size, frame length and mel bins, respectively. The Stem Block is two 2D convolution layers followed by a 2×2 average pooling layer. There are 4 Residual Blocks in the network. Strides for Residual Block 1 to 4 are 1, 2, 2, 1, respectively. After Residual Block 1, an 1×2 average pooling layer is appended to decrease computation. Soft-PS is applied at the outputs of Stem Block and Residual Block 1 to 3.

In Bi-GRU, input size and hidden size are 512. After FC layers, SED and DOA predictions are outputted. The SED predictions bigger than a threshold indicates the event is activate. The DOA predictions are counted under the activate events segement.

To capture time-frequency information efficiently, SA, CBAM and CA are added to Residual Blocks. For SA Residual Block, SplAtConv2d[1] is applied. Cardinals and splits of SplAtConv2d in this report are set to 1 and 2, respectively. The SA Residual Block is shown in Fig. 3(a). CBAM Residual Block is shown in Fig. 3(b). Ratio is set as 8 in channel attention, and kernel size is set as 7 in spatial attention. For the CA Residual Block, the CoordAtt[2] is applied as shown in Fig. 3(c).

3. EXPERIMENT

In this section, dataset of DCASE 2022 Task 3 are firstly described. Then, hyperparameters and training procedure in this report are detailed. Finally, the performance on the test set is presented.

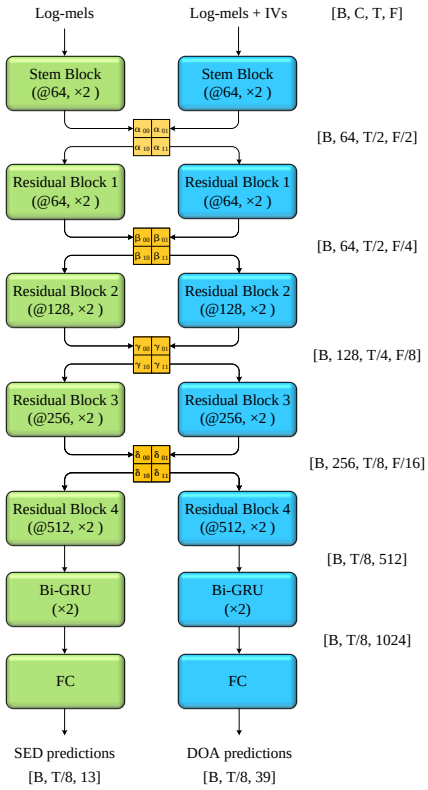

Figure 2: Network architecture

3.1. Dataset

In the dataset of DCASE 2022 Task 3, both synthetic and real recordings are available for development, and performance of the system is tested on the real recordings. 1200 audios with 60sec duration are available in the synthetic set. The duration of real recordings ranges from 30sec to 6min. There are 67 and 54 real recordings in the training set and test set, respectively. And there are 13 target sound event classes. The results in this report are tested on test set, and all of the development set are used to train models.

3.2. Hyperparameters and training procedure

In the baseline system[3], the recordings are segmented by 5sec without overlap. To increase training materials, recordings are segmented by 1sec and their features are extracted. When training the models, the input features are obtained by concatenating 8 adjacent 1sec segments. The hop length is 2sec for synthetic recordings, and 1sec for real recordings. Sample rate, hop length, window length, FFT points, window type and mel bins for input features are 24kHz, 300 samples, 512 samples, 512 samples, Hann window, 128 bins, respectively. Adam optimizer is applied to train SED and DOA branch. Models are trained for 60 epoches. In 1st epoch, learing rate increases from $1e^{-6}$ to $1e^{-3}$. From 2nd epoch to 40th epoch, it decreases from $1e^{-3}$ to $1e^{-5}$. And it holds on $1e{-}5$ over last

[1] https://github.com/zhanghang1989/ResNeSt

[2] https://github.com/Andrew-Qibin/CoordAttention/

[3] https://github.com/sharathadavanne/seld-dcase2022

Table 1: Performances of attention based models on test split

System	$ER_{\leq 20^\circ}$	$F_{\leq 20^\circ}$ (micro)	$F_{\leq 20^\circ}$ (macro)	LE_{CD}	LR_{CD}
Bseline	0.71	36%	21%	29.3°	46%
SA-based CRNN	**0.44**	**66%**	**58%**	**12.9°**	**68%**
CBAM-based CRNN	0.47	64%	52%	14.4°	64%
CA-based CRNN	0.46	65%	55%	14.0°	66%
Attention-based CRNN	0.46	66%	56%	13.7°	67%

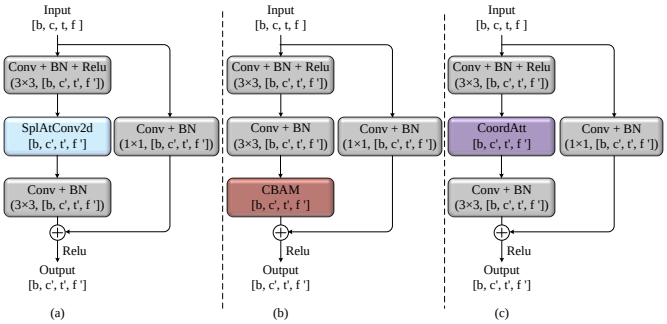

Figure 3: Attention based Residual Block

ten epoches. Models are averaged from checkpoints over last ten epoches. Focal BCE Loss and MSE loss weighted by 0.3 and 0.7 are used to train SED branch and DOA branch. Activation and absence of events are binarized with a threshold of 0.3.

3.3. Inference procedure and metrics

During inference, recordings are cut into 1sec segments. Similar to training stage, window length and hop length of input features are 8sec and 2sec. Then, predictions of overlapping slices are averaged.

The metrics of ER and F for SED are proposed in [11]. The metrics in this report are $ER_{\leq 20^\circ}$, $F_{\leq 20^\circ}$, LE_{CD} and LR_{CD}, which are introduced in [12]. All metrics are computed in 1sec intervals without overlap.

3.4. Results

The results of ensemble of attention based models are shown in table1. The performance of proposed systems are better than the baseline model.

In Table 1, the SA-based CRNN is ensembled by 8 models. The CBAM-based CRNN is ensembled by 7 models. The CA-based CRNN is ensembled by 8 models. And for Attention-based CRNN, it is ensembled by 3 models from SA-based CRNN, 3 models from CBAM-based CRNN, and 3 models from CA-based CRNN. The rank of their performance is SA-based CRNN >Attention-based CRNN >CA-based CRNN >CBAM-based CRNN.

4. CONCLUSION

In this report, SA-based, CBAM-based and CA-based CRNN are applied to solve SELD. Angle noise and mini-batch time-frequency noise are introduced to generalize models. SA-based CRNN achieves the best results in our network framework.

5. ACKNOWLEDGMENT

Thanks to Thi Ngoc Tho Nguyen and Yin Cao for sharing their source codes on github. It has been of a great help to know more details on how to train a SELD model.

6. REFERENCES

[1] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li and A. Smola. "Resnest: Split-attention networks," in *Proc. of IEEE/CVF CVPR*, 2022, pp. 2736–2746

[2] S. Woo , J. Park , J. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. of ECCV*, 2018, pp. 3–19

[3] Q. Hou, D. Zhou and J. Feng, "Coordinate attention for efficient mobile Network design," in *Proc. of IEEE/CVF CVPR*, 2021, pp. 13713–13722

[4] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang and M. D. Plumbley, "An Improved Event-Independent Network for Polyphonic Sound Event Localization and Detection," in *Proc. of IEEE ICASSP*, 2021, pp. 885–889.

[5] 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, pp. 34–48, 2019.

[6] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, "First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation," in *Proc. of DCASE Workshop*, 2019, pp. 154–158.

[7] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. of AAAI CAI*, 2020, pp. 13001–13008.

[8] 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.

[9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[10] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing.*, vol. 28, pp. 2880–2894, 2020.

[11] M. Annamaria, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” *Applied Sciences.*, 6.6(2016):162.

[12] 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, 2021.