GLFE: GLOBAL-LOCAL FUSION ENHANCEMENT FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

Zhengyu Chen

Shanghai University
czhengyu@shu.edu.cn

Qinghua Huang

Shanghai University
qinghua@shu.edu.cn

ABSTRACT

Sound event localization and detection (SELD), as a combination of sound event detection (SED) task and direction of arrival (DOA) estimation task, aims at detecting the different sound events and obtaining their corresponding localization information simultaneously. The more outperforming systems are required to be applied into the more complex acoustic environments. In this paper, our method called global-local fusion enhancement (GLFE) is presented for Detection and Classification of Acoustic Scenes and Events (DCASE) 2022 challenge task 3: Sound Event Localization and Detection Evaluated in Real Spatial Sound Scenes. It could be regarded as a convolution enhancement method. Firstly, the multiple feature cross fusion (MFCF) based on different local receptive fields is proposed. Considering the diversity of real sound events, self-attention network (SANet) integrating global information to local feature is introduced to help the system obtain more efficient information. Further, the skip fusion enhancement (SFE) is explored to fuse the features of different levels by the skip-connection in order to improve feature representation. On Sony-TAu Realistic Spatial Soundscapes 2022 (STRSS22) development dataset, the proposed system shows the significant improvement compared with the baseline system. Series of experiences are implemented only on the first-order Ambisonics (FOA) dataset.

Index Terms— Sound event localization and detection, global-local fusion, self-attention, skip fusion enhancement

1. INTRODUCTION

Sound event localization and detection (SELD), a joint problem involving two areas which are sound event detection (SED) and direction and arrival (DOA), could get the classes of sound events in some time segment and their corresponding elevation and azimuth in 3D space. As an intelligent system, SELD has enormous applications in many areas, such as video surveillance [1-2], robotics and autonomous driving [3-4].

Before deep learning, SED is implemented based on classic methods, such as gaussian mixture model (GMM), support vector machine (SVM), and non-negative matrix decomposition (NMF). DOA is estimated by parameter-based methods, such as multiple signal classification (MUSIC) and estimating signal parameter variational invariance techniques (ESPRIT). Because of the robustness of neural network facing reverberation and noises, deep learning (DL)-based methods are gradually proposed.

The SELD appeared in DCASE 2019. Later, the acoustic complexity of datasets is improved year by year, which varies from the static and single source, moving and polyphony sources, directional interference to real spatial scenes, from 2019 to 2022. Inspired by this, the outperforming models and methods are gradually proposed to boost the robustness for complex acoustic scenes. Adavanne et al [5] proposed SELD-Net which used the convolution neural network (CNN) for high-level space feature extraction and recurrent neural network (RNN) to model sequence context relationship. Many modifications were based on SELD-Net. The intensity vector [6] for FOA and SALSA-Lite [7] for MIC were proposed to extract more efficient manual features. The gated linear unit (GLU) [8] and squeeze-excitation net (SENet) [9] were proposed to improve the ability of feature extraction. TCN [10], Conformer [11] and multi-head self-attention (MHSA) [12] were applied instead of traditional RNN structure. ACCDOA [13] representation, which used a unified regression vector loss instead of a weight sum of BCE and MSE, was proposed to convert double outputs into a single output. Later, multi-ACCDOA [14] was proposed to detect the simultaneous events of the same class.

In this paper, we propose a convolution enhancement method called global-local fusion enhancement (GLFE). It first extracts multiple features by different kernel sizes and fuses them by an efficient method called multiple feature cross fusion (MFCF). Moreover, in order to improve richer information, SANet is adopted to boost the feature representation and further skip connections are used to extract and fuse the features from each convolution block, which is called skip fusion enhancement (SFE). Series of experiences are implemented on FOA development dataset and perform better than the baseline method.

This paper is organized as follows: Section 2 introduces the baseline network and the proposed one. Experiments and discussion are shown in Section 3. Finally, the conclusion is shown in Section 4.

2. METHODS

2.1. Baseline Network

The baseline is adopted based on [14]. The input feature is multi-channel intensity vector (IV) for FOA dataset or SALSA-Lite for MIC dataset. The manual features are first fed into three successive convolution blocks to extract high-level features, and the max pooling and dropout are put between each convolution block to decrease overfitting. Moreover, the extracted features are fed into GRU modeling sequence context information. Finally, the ACCDOA representation is obtained by two fully connected layers.

2.2. Multiple Feature Cross Enhancement

The proposed SELD framework is represented in figure 1. Considering the density diversity of real acoustic scenes, the normal CNN could not be robust for feature extraction. Based on this, we modified the feature extractor by the proposed GLFE. In the framework, we first fuse multiple features inter-sectionally which have different receptive fields and are obtained by different kernel sizes. The framework of MFCF is shown in figure 2, which integrates global and local information simultaneously. The SAnet is introduced for extracting global information and then the skip connection is applied, which fuses the output of each block to obtain richer information by SFE blocks. The detail of SFE block is shown in figure 3. It is worth noting that the kernel size of the pooling block is different for each SFE block in order to fuse correctly.

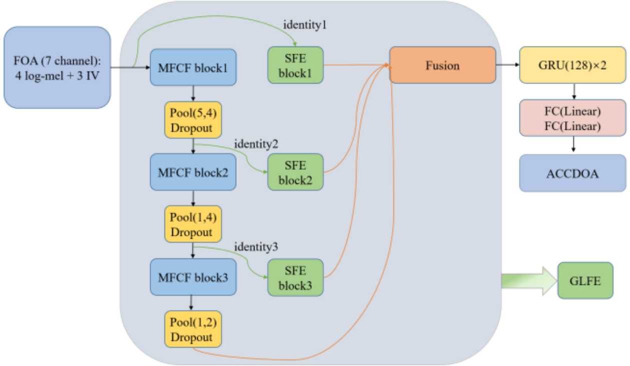

Figure 1: SELD framework proposed in this work. The shaded area is the proposed GLFE module instead of convolution block in the baseline network. The fusion adopts concatenation operation along channel dimension.

Figure 2 give the details about MFCF. We use the Conv2D whose kernel sizes are 3, 5, 1 to extract multiple features with different receptive fields. Further we fuses them inter-sectionally to boost the feature representation. Moremover, the self-attention (SAblock) is introduced in order to get richer information facing with real acoustic scenes. Finally, the outputs of LOCAL and SAblock are fused as a output of one MFCF block. The conv2D (3×3) and (5×5) use the conv_BN_SiLU model and conv2D (1×1) uses the conv_BN constructure.

The SFE block are shown in figure 3. The kernel sizes of pooling blocks are different in order to obtain the same feature size before the fusion operation. The SFE block 1 needs three successive pooling operations whose kernel sizes are (5, 4), (1, 4), (1, 2), respectively. The SFE block 2 needs two successive pooling operations whose kernel sizes are (1, 4) and (1, 2) , respectively. The SFE block 3 needs one pooling operation whose kernel sizes are (1, 2).

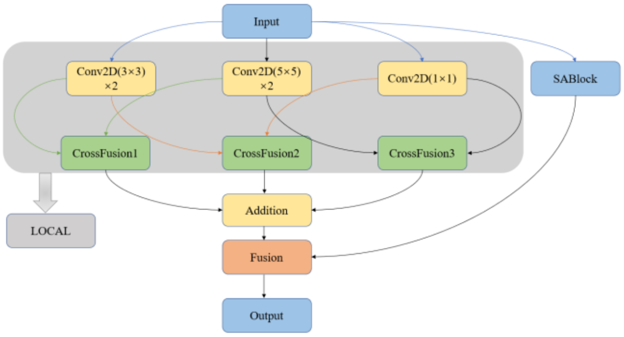

Figure 2: The specific structure for MFCF block. The crossfusion adopts the addition operation and the fusion uses the concatenation operation.

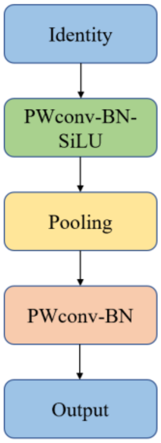

Figure 3: The structure of SFE block. PWConv is point-wise convolution.

3. EXPERIMENTS AND DISCUSSION

3.1. Dataset

FOA data formats are used in the Sony-TAu Realistic Spatial Soundscapes 2022 (STARS22) which provides 67 recordings for training and 54 recordings for testing on the development dataset. Considering the diversity of real sound density, we adopt the external dataset to train, i.e., 750 recordings in FSD50K which include 1200 recordings the official provides.

3.2. Training Procedure and Evaluation Metrics

In order to have a fair comparison between the baseline and proposed framework, the hyper-parameter setting almost is not changed, except for the dropout is modified into 0.1 and the batch size is set as 50.

The evaluation metrics are still ER_{20° and F for SED, LE_{CD} and LR_{CD} for DOA, and the comprehensive metric SELDscore. However, this year the task performs macro-averaging mode, which computes the metrics for each class and then averages them along the class.

3.3. Results and Discussion

We compare the results based on the baseline network and proposed GLFE. Table 1 gives the detail values to evaluate them. The results show better performance obtained by the proposed method than the baseline. Especially the F score is improved 3%, LR_{CD} is improved 4%, and LE_{CD} decreases 13.42%. Figure 4 shows the change of SELDscore (SELD) for each class. The whole trend shows that the proposed framework has better results.

Table 1: The performance comparison for different methods on development dataset. The ED means that the external dataset is used.

Method	$ER_{20°}$	F(%)	LE_{CD}	LR_{CD}(%)	SELD
Base	0.79	11%	65.72	25%	0.70
BaseED	0.72	24%	40.1	44%	0.57
Pro-ED	**0.71**	**27%**	**26.68**	**48%**	**0.53**

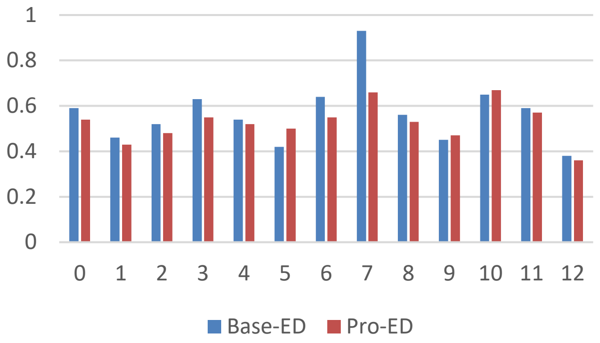

Figure 4: The SELDscore comparison of the baseline system and proposed system for each class. The number of horizontal axis is the class number. The vertical axis represents the SELDscore.

4. CONCLUSION

We propose the global-local fusion enhancement as a convolution enhancement method to boost the feature representation of SELD system. First, we fuse the features crossover with different receptive fields obtained by different convolution kernel sizes. The self-attention block is integrated with local features to help get richer information. Moreover, the SFE is proposed in our framework to fuse the features with different levels. In order to train better, the external dataset is adopted. The experimental results compared with the baseline shows better performance.

5. REFERENCES

[1] M. Crocco, M. Cristani, A. Trucco and V. Murino, "Audio surveillance: A systematic review", ACM Comput. Surv., vol. 48, 2016.

[2] C. Grobler, C. Kruger, B. Silva and G. Hancke, "Sound based localization and identification in industrial environments", Proc. 43rd Anuu. Conf. IEEE Ind. Electron. Soc., pp. 6119-6124, 2017.

[3] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information", Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp. 405-409, 2016.

[4] W. He, P. Motlicek and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization", Proc. Int. Conf. Robot. Autom., pp. 74-79, 2018

[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, 2018.

[6] P .-A. Grumiaux, S. Kitic, L. Girin, and A. Guérin, “Improved feature extraction for CRNN-based multiple sound source localization,” in Proc. Europ. Signal Process. Conf. (EUSIPCO), Dublin, Ireland, 2021.

[7] T. N. Tho Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan and W. -S. Gan, "SALSA-Lite: A Fast and Effective Feature for Polyphonic Sound Event Localization and Detection with Microphone Arrays," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 716-720, doi: 10.1109/ICASSP43922.2022.9746132.

[8] T. Komatsu, M. Togami and T. Takahashi, "Sound Event Localization and Detection Using Convolutional Recurrent Neural Networks and Gated Linear Units," 2020 28th European Signal Processing Conference (EUSIPCO), 2021, pp. 41-45, doi: 10.23919/Eusipco47968.2020.9287372.

[9] Naranjo-Alcazar J, Perez-Castanos S, Ferrandis J, et al. Sound Event Localization and Detection using Squeeze-Excitation Residual CNNs[J]. arXiv preprint arXiv:2006.14436, 2020.

[10] K. Guirguis, C. Schorn, A. Guntoro, S. Abdulatif and B. Yang, "SELD-TCN: Sound Event Localization & Detection via Temporal Convolutional Networks," 2020 28th European Signal Processing Conference (EUSIPCO), 2021, pp. 16-20, doi: 10.23919/Eusipco47968.2020.9287716.

[11] Zhang Y, Wang S, Li Z, et al. Data Augmentation and Class-Based Ensembled CNN-Conformer Networks for Sound Event Localization and Detection[R]. Technical Report of DCASE Challenge. 2021. Available online: http://dcase.community/documents/challenge2021/technical_reports/DCASE2021_Zhang_67_t3. pdf (accessed on 8 May 2022), 2021.

[12] Emmanuel P, Parrish N, Horton M. Multi-scale network for sound event localization and detection[R]. Tech. report of DCASE Challenge, 2021.

[13] 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," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021,pp.915-919,doi: 10.1109/ICASSP39728.2021.9413609.

[14] 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," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 316-320, doi: 10.1109/ICASSP43922.2022.9746384.