SOUND EVENT LOCALIZATION AND DETECTION USING CROSS-MODAL ATTENTION AND PARAMETER SHARING FOR DCASE2021 CHALLENGE

Technical Report

Sang-Hoon Lee *, Jung-Wook Hwang, Sang-Buem Seo, Hyung-Min Park
Dept. of Electronic Engineering, Sogang University
35 Baekbeom-ro, Mapo-gu, Seoul 04107, South Korea
{shleee47, hju0518, sbseo, hpark}@sogang.ac.kr

ABSTRACT

In this report, we present our model for DCASE2021 Challenge Task3: Sound Event Localization and Detection (SELD) with Directional Interference. The model learns sound event detection (SED) and direction-of-arrival (DoA) at once by multi-task learning for the SELD task. When learning the model, general features for both SED and DoA prediction are extracted by using the parameter-sharing strategy at the feature level of SED and DoA. In addition, the output is estimated by adding an attention layer based on cross-modal attention (CMA) in the transformer decoder so that the system can efficiently learn associations between SED and DoA features. Furthermore, three different prediction rules are presented for fully connected (FC) networks to provide SED and DoA results. Experiment has been conducted on the TAU-NIGENS Spatial Sound Events 2021 dataset, and to produce more learning data, the data was augmented by the mixup to sum up weighted audio clips and the channel rotation to change the location information of the sound source by rotating input channels, in addition to SpecAugment. Experimental results showed that our method provided significantly improved performance than the baseline method.

Index Terms— DCASE2021 task3, sound event localization and detection, cross-modality, transformer, attention, parameter-sharing, data-augmentation

1. INTRODUCTION

The Sound Event Localization and Detection (SELD) is a complex task to detect individual sound events in a class with sound event detection (SED) and to estimate their positions for directional events that do not belong to ambient noise with direction-of-arrival (DoA) estimation. For example, it can be used in situations that occur inside a house, such as a baby crying or a window broken, or it can be used in situations that occur outside, such as a car horn on the road, collapse of a building, and a disaster relief call. In a practical situation, sound of an event is transmitted to microphones from the corresponding source at a specific direction. From this fact, it is reasonable to combine detection and localization by estimating the temporal and spatial location of the event. Therefore, it is worthwhile to study SED and DoA together, and the task has recently become a popular topic after DCASE2019 and 2020.

In DCASE2021, the baseline model [1] of the SELD system learns SED and DoA models jointly with commonly extracted features. The joint learning is adopted to avoid the data association problem between the predicted sound events and the estimated DoAs with separated SED and DoA estimation [2]. In the baseline model, features are extracted by convolutional neural network (CNN) followed by recurrent neural network (RNN) for audio input data and commonly used for SED and DoA models. Each model is composed of a two-layer fully connected (FC) network. The SED is performed with a multi-label classification task and estimates which event exists for each frame while the DoA estimate is obtained in three-dimensional coordinates with a multi-output regression task.

If common features are extracted like the baseline model, some features will be trained for SED (or DoA), and then it will be difficult to learn for DoA (or SED). In this report, unlike the baseline model, we obtain each feature sequence for SED or DoA using CNN layers in parallel. However, features for the two models should be associated because the two tasks are directly related. Therefore, we use the parameter-sharing method in [3] that exchanges intermediate features in the CNN layers for the SED and DoA. In addition, the performance of the SED and DoA is further improved by using cross-modal attention (CMA) in transformer decoder layers to learn the fused information.

2. DATASETS

The DCASE2021 Task3 provides two formats of the TAU-NIGENS Spatial Sound Events 2021 dataset: a tetrahedral microphone array one (MIC) and first-order Ambisonics one (FOA). In this paper, we only used the FOA (4-channels, 3-dimensional recordings) data format for the experiments.

Contrary to the DCASE2020 Task3 dataset, the DCASE2021 dataset contains 12 different SED target classes, and up to three overlapping events may occur. In addition, there exist diverse sound events not belonging to any of the target classes which have their own temporal activities from a static or moving source.

* This work was supported by Institute of Information and communications Technology Planning and evaluation (IITP) Grant funded by the Korea government (MSIT) (No. 2019-0-01376, Development of the multi-speaker conversational speech recognition technology).

3. PROPOSED APPROACH

In this paper, we apply CMA for the first time to SELD. Several multimodal tasks such as audio-visual speech recognition [4] or audio-text emotion detection [5] use the CMA in transformer decoder layers to complement features for each modality. However, we rarely see such attempts in multi-task learning.

This task requires simultaneous prediction of the classes and direction of arrivals (DoA) of sound events, where associations between features for SED and features for DoA may be helpful. In this respect, we expect our CMA-based model structure to efficiently learn the associations. We refer to our proposed model as CMA-SELD.

3.1. Features and Data Augmentation

Our proposed model architecture is largely divided into SED and DoA streams. The log-mel spectrogram extracted from input audio data is used as input features for SED, and intensity vectors are additionally extracted as input features for DoA estimation.

We use three data augmentation methods. First, we use the mixup that sums up weighted audio clips [6]. Second, we use the channel rotation by exchanging audios by channel [7]. Third, we apply time-frequency masking to input spectrograms via SpecAugment [8].

3.2. Convolutional Embedding

The embedding process consists of CNN blocks. Although we have also tried embeddings with the LSTM or transformer encoder, nothing performed better than the CNN. Inspired by Yin Cao [3], we perform parameter sharing between features in CNN blocks for DoA and SED. DoA features are calculated by the sum of CNN outputs in the SED and DoA streams weighted by learnable parameters $\mathbf{\Phi_C} = [\Phi_1, \Phi_2 \ldots \Phi_c \ldots, \Phi_C]$ and $\boldsymbol{\theta_C} = [\theta_1, \theta_2 \ldots \theta_c \ldots, \theta_C]$, where C is the number of channels. Similarly, SED features are calculated by the sum of CNN outputs in the DoA and SED streams weighted by learnable parameters $\boldsymbol{\alpha_C} = [\alpha_1, \alpha_2 \ldots \alpha_c \ldots, \alpha_C]$ and $\boldsymbol{\beta_C} = [\beta_1, \beta_2 \ldots \beta_c \ldots, \beta_C]$. The features with parameter sharing are expressed as follows:

$$DoA = \theta * DoA_{CNN} + \Phi * SED_{CNN}, \quad (1)$$

$$SED = \alpha * DoA_{CNN} + \beta * SED_{CNN}, \quad (2)$$

where DoA and SED denote DoA and SED features with parameter sharing, respectively, and DoA_{CNN} and SED_{CNN} represent CNN outputs in the SED and DoA streams, respectively.

3.3. Cross-Modal Attention

Output embeddings are fed to transformer decoder layers. After self-attention, the CMA module in the DoA stream (CMA-1) takes the DoA embedding sequence as a query vector and the SED embedding sequence as a key and value vector for the multi-head scale dot production attention. Likewise, the CMA module in the SED stream (CMA-2) takes the SED embedding sequence as a query vector and the DoA embedding sequence as a key and value vector for the multi-head scale dot production attention. The CMA is followed by a feed-forward layer. The feed-forward layer and self-attention are omitted in Fig. 1. The number of layers in the transformer decoder was set to two. In addition to the convolutional embedding, this process will play a major role in obtaining the association between SED and DoA feature information prior to the final detection.

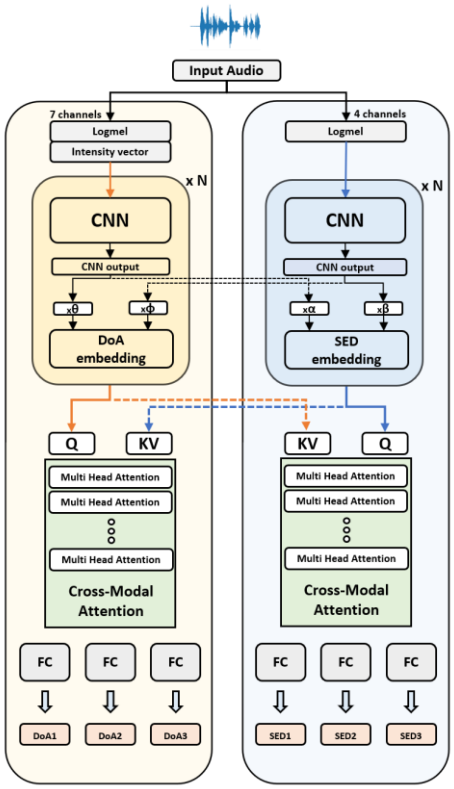


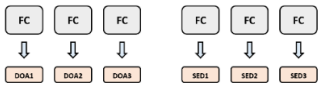
Figure 1: Overall CMA-SELD architecture.

3.4. Fully Connected Networks

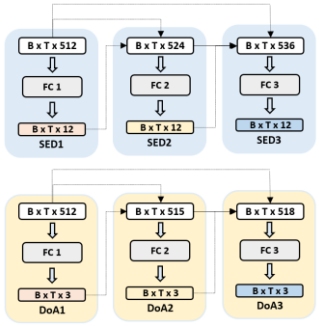
The DCASE2021 Task3 can have up to three overlapping events. Three SED outputs and three DoA outputs are obtained through an FC network for each output. We present three different prediction rules for the FC networks as shown in Fig. 2.

First, Simple Parallel Prediction simply passes the output of the last transformer layer of the CMA-1 or CMA-2 into each FC network in parallel. As a disadvantage of this approach, the results of three FC networks may be duplicated. For example, when Event A, B, and C occur simultaneously at a particular frame, all the three FC networks may accidentally predict A.

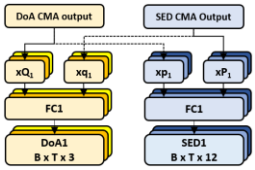
To avoid this, Serial Prediction is attempted. The outputs of the previous FC networks are concatenated to the input of the current FC network to predict an event different from the previous one. In the DoA stream, the first, second, and last FC networks take (512), (512+3), and (512+3+3) dimensions, respectively, whereas the first, second, and last FC networks take (512), (512+12), and (512+12+12) dimensions in the SED stream, respectively.



(a) Simple Parallel Prediction



(b) Serial Prediction



(c) Prediction with Parameter Sharing

Figure 2: Three different prediction rules for FC networks.

Table 1: Results of our proposed model. '1' and '2' after Simple Parallel denotes the number of transformer decoder layers

Methods	$ER_{<20^\circ}$	$F_{<20^\circ}$	LE_{CD}	LR_{CD}
Baseline	0.69	33.9%	24.1°	43.9%
Simple Parallel-1	0.48	59.9%	**14.3°**	73.2%
Simple Parallel-2	**0.46**	**60.9%**	14.4°	73.3%
Serial	0.48	59.4%	15.1°	**73.4%**
Parameter Sharing	0.49	59.4%	15.1°	**73.4%**

The last rule is Prediction with Parameter Sharing, where the parameter sharing is performed for inputs to the corresponding FC networks in the DoA and SED streams. We expect the parameter sharing to make predictions of DoA and SED for an identical event.

4. RESULTS

The results of our proposed model with the three prediction rules introduced in Section 3 are shown in Table 1. Additionally, we also show the results of the Simple Parallel Prediction with a one-layer transformer decoder. Introducing CMA and parameter sharing in the SELD task, all the experimented results for the proposed method outperformed the baseline performance significantly. The highest performance was indicated in bold.

5. CONCLUSION

In this paper, we proposed a method based on CMA and parameter sharing to simultaneously detect and localize sound events. The CNN-based encoder with parameter sharing exchanges intermediate features in the CNN layers for the SED and DoA, and the CMA-based decoder efficiently outputs three predictions to detect up to three overlapping events in the DCASE2021 Task3. Furthermore, three different prediction rules for the FC networks were considered. In particular, a key and value vector for the CMA was given from the other stream to efficiently learn associations between SED and DoA features. Experimental results show the proposed system outperformed the baseline methods significantly.

As a future work, we will study on prediction rules to induce non-independent and reliable results between the FC networks. For further improvements, we plan on developing the network to output the number of events occurring concurrently as well as the classes of events to deal with unlimited events simultaneously occurring in real-world situations.

6. ACKNOWLEDGMENT

Yin Cao's research [3] was of great help to this study. We would like to thank Yin Cao for sharing the research on the parameter sharing method on SELD.

7. REFERENCES

[1] S. Adavanne, A. Politis, J. Nikunen, and Tuomas Virtanen, "Sound event localization and detection of overlapping sources using convolutional re-current neural network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, pp. 34-48, 2018.

[2] T. Butko, F. González Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: Online implementation in a smart-room," in *Proc. EUSIPCO*, 2011, pp. 1317-1321.

[3] 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. IEEE ICASSP*, 2021, pp. 885-889.

[4] Y.-H. Lee, D.-W. Jang, J.-B. Kim, R.-H. Park, and H.-M. Park, "Audio–visual speech recognition based on dual cross-modality attentions with the transformer model," *Applied Sciences*, vol. 10, 2020.

[5] A. Khare, S. Parthasarathy, and S. Sundaram, "Self-supervised learning with cross-modal transformers for emotion recognition," in *Proc. IEEE SLT*, 2021, pp. 381-388.

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

[7] F. Ronchini, D. Arteaga, and A. Perez-Lopez, "Sound event localization and detection based on CRNN using rectangular filters and channel rotation data augmentation," *arXiv:2010.06422*, 2020.

[8] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613-2617.