Sound Event Detection Using Attention and Aggregation-Based Feature Pyramid Network

Ji Won Kim
AI Graduate School
Gwangju Institute of Science and Technology (GIST)
Gwangju, Korea
jiwon.kim@gm.gist.ac.kr

Geon Woo Lee
AI Graduate School
Gwangju Institute of Science and Technology (GIST)
Gwangju, Korea
geonwoo0801@gist.ac.kr

Hong Kook Kim*
School of EECS,
AI Graduate School
Gwangju Institute of Science and Technology (GIST)
Gwangju, Korea
hongkook@gist.ac.kr

Nam Kyun Kim
Automotive Electronics R&D Center
Korea Automotive Technology Institute
Gwangju, Korea
kimnk@katech.re.kr

***Abstract*—This paper proposes a sound event detection (SED) model using an EfficientNet-B2 and an attention and aggregation-based feature pyramid network (A2-FPN). In particular, the EfficientNet-B2 is first obtained from the pretrained model on the basis of the pretraining, sampling, labeling, and aggregation (PSLA) framework. Then, the A2-FPN module is applied to the outputs of the layers of the EfficientNet-B2 to deal with the different time and frequency resolutions from acoustic features. The aggregated feature map from the A2-FPN module is used as input features to two bidirectional gated recurrent unit layers. Specifically, the proposed A2-FPN-based SED model is trained by the mean-teacher approach to utilize weakly labeled and unlabeled data. Finally, the proposed A2-FPN-based SED model is applied to the detection and classification of acoustic scenes and events (DCASE) 2021 Challenge Task 4. Consequently, it is shown that the polyphonic sound event detection score (PSDS) 1 and 2 of the proposed A2-FPN-based SED model are the higher of 0.03 and 0.172, respectively, than those of the DCASE 2021 Challenge Task 4 baseline.**

Keywords*—*sound event detection, attention and aggregation-based feature pyramid network, EfficientNet, DCASE 2021 Challenge Task 4

I. Introduction

Sound event detection (SED) aims to detect sound events from acoustic signals and classify them into individual sound event categories with timestamps. SED has been widely used to support sound sensing applications, such as wildlife monitoring, equipment monitoring, audio captioning, and etc. [1]. Recently, neural network-based SED models have been developed rapidly. Among them, SED models, which aggregate features with different resolutions according to the time and frequency domains, have been successfully performed [2]. By doing this, it is expected that these SED models should be better than single-scale resolution-based SED models, when repeated sound events occur, such as frequent barking or alarm bell ringing, or when a wide range of frequencies is occupied, such as during vacuum cleaning.

To deal with multi-scale resolution, a residual network (ResNet) with the attention and aggregation-based feature pyramid network (A2-FPN) module has been proposed for image sematic segmentation [3]. The A2-FPN module comprises an attention and aggregation (A2) module and a feature pyramid network (FPN) module. The feature maps generated from each convolutional layer in the ResNet are combined by the FPN and A2 module procedures. In other words, the feature map from a deeper layer can represent higher level features with a lower resolution, compared to that from a shallower layer. Thus, combining the feature maps from the shallower to deeper layers can represent different resolutions and levels from the input features.

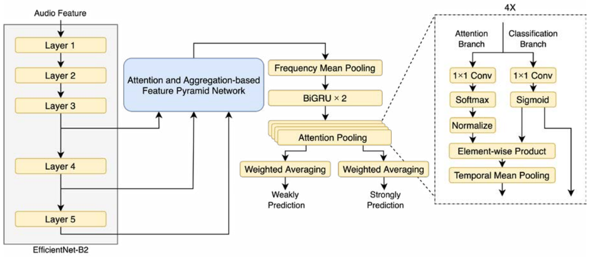

Fig. 1. Network architecture of the proposed SED model using the EfficientNet-B2 and A2-FPN module.

In this paper, we propose an SED model using an EfficientNet-B2 and an attention and A2-FPN. Then, the proposed SED model is evaluated on the detection and classification of acoustic scenes and events (DCASE) 2021 Challenge Task 4.

II. Proposed Sound Event Detection Model

A. Network Architecture

Fig. 1 shows the network architecture of the proposed A2-FPN-based SED model using an EfficientNet-B2. First of all, a sequence of 128-dimensional mel-spectra is obtained by applying a short-time Fourier transform to each audio file. Then, these input features are applied to the EfficientNet-B2 that is priory obtained from the pretrained model on the basis of the pretraining, sampling, labeling, and aggregation (PSLA) framework. The reason why we used the PSLA framework is that it was proved to show good performance in audio tagging tasks [4]. In the EfficientNet-B2, the frequency and time resolutions are reduced by half after passing through each layer, except for the last layer, resulting in different resolutions according to the layers in EfficientNet-B2.

Next, the A2-FPN module is applied to the EfficientNet-B2 to aggregate feature maps with different resolutions. After that, the aggregated feature map from the A2-FPN module is processed by a frequency mean pooling layer, bidirectional gated recurrent unit (BiGRU) layers, an attention pooling layer, and weighted averaging layers sequentially. In particular, the BiGRU layers are used to acquire sequential information. Particularly, the attention pooling layer has four independent

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00963, Localization Technology Development on Spoken Language Synthesis and Translation of OTT Media Contents).

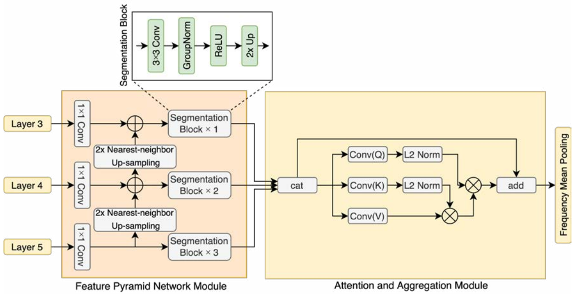

Fig. 2. Network architecture of the A2-FPN module in the proposed SED model.

heads, where each head comprises an attention branch and a classification branch [4], as shown in Fig. 1. The attention branch is a 1×1 convolution layer with a softmax function, followed by a normalization layer, while the classification branch is a 1×1 convolution layer with a sigmoid function. The output of each branch is subjected to an element-wise product, followed by temporal mean pooling, which gives the probability for weakly prediction. This weak prediction represents for sound event categories without timestamps. In contrast, the output from the classification branch is also a logit function representing for event categories and timestamps. Finally, each of the logits and probabilities is scaled using the same learnable parameters and summed for each prediction.

B. Attention and Aggregation-Based Feature Pyramid Network Module

Fig. 2 illustrates the network architecture of the A2-FPN module in the proposed SED model. At first, the layer 3, 4, and 5 of the EfficientNet-B2 provides the output feature maps whose shapes are (48×16×132), (120×8×66), and (352×4×33), respectively. Note here that the shape of (x × y × z) means (channel × frequency × frame). Then, each feature map is applied with each 1×1 convolution layer with 48 outputs. Next, to match the resolution between a lower and its upper layer, a nearest-neighbor up-sampling procedure is applied to the lower layer. In other words, two up-sampling procedures are applied to EfficientNet-B2 layers 3–4 and 4–5, and each of the up-sampled feature maps is added to each feature map of the upper layer, as shown in the figure. These added feature maps are then used as the input features of the segmentation blocks, where different number of the segmentation blocks is employed from top to bottom layers. Note that we do not use the layer 6 of the EfficientNet-B2 because there is no resolution variation in the layer.

Subsequently, the outputs of the segmentation blocks are aggregated by channel-wise concatenation, and then they are further processed by linear self-attention [3], followed by an add operation. Finally, the output of the A2 module is inputted into a frequency mean pooling layer.

C. Semi-Supervised Learning Using Mean-Teacher

To train the proposed A2-FPN-based SED model on the DCASE 2021 Challenge Task 4, the mean-teacher approach is used. This is because the mean-teacher approach is a kind of semi-supervised learning methods so that it can deal with unlabeled and weakly labeled data that are included in the training dataset [5]. In other words, the mean-teacher approach is able to train the proposed SED model by comparing mean-squared errors between the output of the student and teacher models. To reduce generalization error, several augmentation techniques such as time-frequency shift, time mask, mix-up, and filter-augmentation are applied to the training data [5].

TABLE I. PERFORMANCE COMPARISON OF THE BASELINE, EFFICIENTNET-B2-BASED, AND PROPOSED AF-FPN-BASED SED MODEL ON THE EVALUATION DATASET OF DCASE 2021 CHALLENEGE TASK 4

MODEL	PSDS 1	PSDS 2
DCASE 2021 Baseline [6]	0.407	0.627
EfficientNet-B2	0.282	0.667
EfficientNet-B2 w/ A2-FPN (Proposed)	0.437	0.799

III. PERFORMANCE EVALUATION

The proposed A2-FPN-based SED model was applied to the DCASE 2021 Challenge Task4 [6]. The training and evaluation data were stereo audio sampled at 48 kHz, and the left channel signal of each stereo audio was resampled with 16 kHz. Then, the resampled audio was segmented into frames, where each frame had 2,048 samples with 160 hop length. Next, a 128-dimensional mel-spectrum was extracted by applying the 2,048-point fast Fourier transform to each frame.

Next, the proposed SED model was evaluated on the public evaluation dataset using the polyphonic sound detection score (PSDS) 1 and 2 defined in DCASE 2021 Challenge Task 4 [7]. Table I compares the performances of the DCASE baseline [6], the EfficientNet-B2-based, and proposed A2-FPN-based SED model on the public evaluation dataset. As shown in the table, the EfficientNet-B2-based SED model achieved higher PSDS 2 but lower PSDS 1 than the baseline. However, the proposed SED model, which was the EfficientNet-B2 with the A2-FPN module, showed the higher PSDS 1 and 2 by 0.03 and 0.172, respectively, than the baseline, and also it provided significantly better than the EfficientNet-B2-based SED model.

IV. CONCLUSION

In this paper, we proposed an A2-FPN-based SED model using the EfficientNet-B2, and applied it to the DCASE 2021 Challenge Task 4. It was shown from the performance evaluation that the proposed SED model significantly improved PSDS 1 and 2, compared to the DCASE 2021 baseline and the EfficientNet-B2-based SED model. This implied that the A2-FPN module helped the SED model detect sound events with different resolutions and levels.

REFERENCES

[1] T. K. Chan and C. S. Chin, "A comprehensive review of polyphonic sound event detection," IEEE Access, vol. 8, pp. 103339–103373, 2020.

[2] D. De Benito-Gorrón, D. Ramos, and D. T. Toledano, "A multi-resolution CRNN-based approach for semi-supervised sound event detection in DCASE 2020 challenge," IEEE Access, vol. 9, pp. 89029–89042, 2021.

[3] R. Li, L. Wang, C. Zhang, C. Duan, and S. Zheng, "A2-FPN for semantic segmentation of fine-resolution remotely sensed images," Int. J. Remote Sens., vol. 43, no. 3, pp. 1131–1155, 2022.

[4] Y. Gong, Y.-A. Chung, and J. Glass, "PSLA: Improving audio tagging with pretraining, sampling, labeling, and aggregation," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 29, pp. 3292–3306, 2021.

[5] N. K. Kim and H. K. Kim, "Polyphonic sound event detection based on residual convolutional recurrent neural network with semi-supervised loss function," IEEE Access, vol. 9, pp. 7564–7575, 2021.

[6] Sound Event Detection and Separation in Domestic Environments, https://dcase.community/challenge2021/task-sound-event-detection-and-separation-in-domestic-environments.

[7] G. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic´, "A framework for the robust evaluation of sound event detection," in Proc. ICASSP, 2020, pp. 61–65.