SEMI-SUPERVISED LEARNING-BASED SOUND EVENT DETECTION USING FREQUENCY-CHANNEL-WISE SELECTIVE KERNEL FOR DCASE CHALLENGE 2022 TASK 4

Technical Report

Ji Won Kim[1], *Geon Woo Lee*[1], *Hong Kook Kim*[1,2,*], *Yeon Sik Seo*[3], *and Il Hoon Song*[3]

[1] AI Graduate School, [2] School of Electrical Engineering and Computer Science
Gwangju Institude of Science and Technology, Gwangju 61005, Korea
jiwon.kim@gm.gist.ac.kr, {geonwoo0801, hongkook}@gist.ac.kr
[3] AI Lab., R&D Center, Hanwha Techwin, Seongnam-si, Gyeonggi-do 13488, Korea
{yeonsik.seo, ilhoon}@hanwha.com

ABSTRACT

In this report, we propose a mean-teacher model-based sound event detection (SED) model that uses semi-supervised learning to the labeled data deficiency problem for the DCASE 2022 Challenge Task 4. The mean-teacher model of the proposed SED model is based on a residual convolutional recurrent neural network (RCRNN) architecture, and the residual convolutional blocks in the RCRNN are modified to include the frequency-wise and/or channel-wise selective kernel attention (SKA), which is hereafter referred to as SKA-RCRNN. This enables the RCRNN to have an adaptive receptive field for different lengths of audio. In particular, the proposed SKA-RCRNN-based SED model is first trained on the training dataset, during which it generated pseudo-labeled data for weakly labeled and unlabeled data. Next, the noisy student model, which is also based on SKA-RCRNN, in the second stage is optimized via semi-supervised learning by using strongly labeled and pseudo-labeled data. Finally, several ensemble models are obtained from fivefold cross-validation SED models with various hyper-parameters, and some of them are selected as the submitted models that show higher F1 and polyphonic sound detection scores on the validation dataset of the DCASE 2022 Challenge Task 4 are selected for submission.

Index Terms— Sound event detection, semi-supervised learning, noisy-student model, residual convolutional recurrent neural network (RCRNN), frequency-wise and channel-wise selective kernel attention (SKA)

1. INTRODUCTION

Sound event detection (SED) aims to detect sound events from acoustic signals and classify them into individual sound event categories with timestamps in diverse acoustic environments. SED models have been widely used to support sound-sensing applications, such as wildlife monitoring [1], equipment monitoring [2], and audio captioning [3]. Recently, the availability of large amounts of strongly labeled data that include correct event categories and well-refined timestamps has substantially improved the performance of SED models. However, strongly labeling of data is extremely expensive and might contain human errors, which cause the SED model's performance to deteriorate. Alternatively, weakly labeled data, whose labels only include the sound event categories without any information on the timestamps, can be used for SED modeling in combination with a limited amount of strongly labeled data. Furthermore, unlabeled data, whose labels are timestamps only without any information on sound event categories, could potentially be utilized for model training, as in the DCASE Challenge.

The DCASE 2022 Task 4 is the follow-up to the DCASE 2021 Task 4. Compared to the DCASE 2021 Task 4, additional strongly labeled data are included from AudioSet [4], while the weakly labeled and unlabeled data are identical in both years' tasks. Specifically, it is allowed that the participants can use pretrained models that are trained using an additional dataset. This year, we try to improve the SED model that was submitted to DCASE 2021 Challenge Task 4 by incorporating a selective kernel attention (SKA) to some of the residual convolutional layers in a residual convolutional recurrent neural network (RCRNN). The convolutional layer with a fixed sized kernel focuses on representing local information in time or frequency, which might result in poor SED performance when repeated sound events occur, such as frequent barking or alarm bell ringing, or when a wide range of frequencies is occupied, like during vacuum cleaning.

To remedy this problem, we propose an SKA-RCRNN model for SED, where SKA is implemented via channel-wise selective kernel attention (cwSKA), frequency-wise selective kernel attention (fwSKA), or both. In addition, two-stage and semi-supervised learning strategies using the mean-teacher model are applied to train the SED model with weakly labeled and unlabeled data. In the first stage, the proposed SKA-RCRNN-based SED model is trained based on a consistency loss function by using the entire training dataset, including strongly labeled, weakly labeled, and unlabeled data, during which it generated pseudo-labeled data for weakly labeled and unlabeled data. Next, the noisy student model, which is also an SKA-RCRNN, in the second stage is optimized on the basis of the self-learning with a semi-supervised loss function by using strongly labeled and pseudo-labeled data.

Following this introduction, Section 2 describes the dataset and the input features of the SED model used in this work. Section 3 proposes the SKA-RCRNN model and learning strategy. Then, Section 4 discusses the experimental results of the proposed SED model on the validation dataset for the DCASE 2022 Task 4. Finally, Section 5 concludes this report.

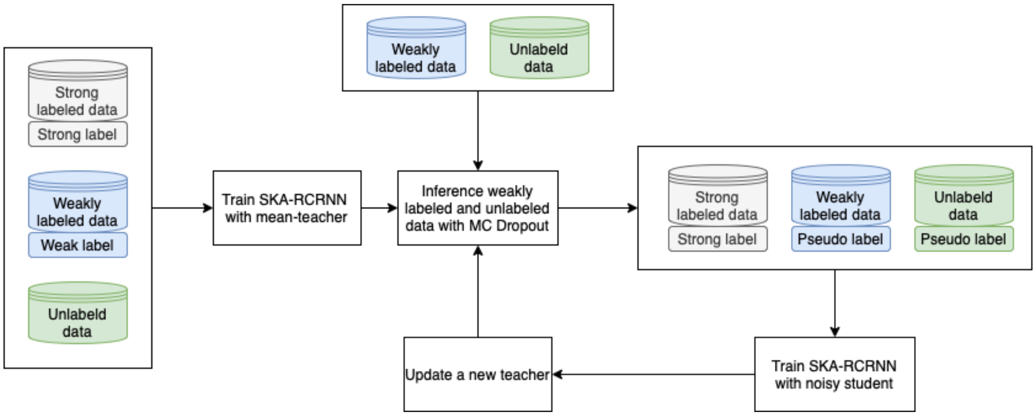

Figure 1: Training procedure of the proposed SKA-RCRNN-based SED model, where a two-stage mean-teacher model composed of SKA-RCRNN was trained by semi-supervised learning.

2. DATASET

The DCASE 2022 Challenge Task 4 consists of four distinct datasets for model training: 1) a weakly labeled training dataset (without timestamps), 2) an unlabeled in-domain training dataset without any labels, 3) a strongly labeled synthetic dataset, and 4) a strongly labeled real dataset. The weakly labeled and unlabeled in-domain training datasets are taken from AudioSet, while the strongly labeled synthetic dataset is generated using the Scaper soundscape synthesis and augmentation library [5]. The weakly labeled training dataset contains 1,578 audio clips with weak annotations only. The unlabeled in-domain training dataset contains 14,412 audio clips without any labels. Finally, the synthetic and real strongly labeled datasets contain 10,000 and 3,470 audio clips, respectively. Each audio clip is sampled at 44.1 kHz with a maximum duration of 10 seconds.

For the given dataset, we process the data for the model input. First, the mono-channel signals are resampled from 44.1 to 16 kHz. Next, each resampled audio signal is segmented into consecutive frames of 2,048 samples with 256 samples of hop length. Then, a 2,048-point fast Fourier transform (FFT), followed by a 128-dimensional mel-filterbank analysis, is performed for each frame. Since each 10-second audio clip is represented by 625 frames, the dimension of the input feature is 625×128. Finally, the extracted mel-spectrogram features are normalized using the global mean and the standard deviation for all of the training audio clips.

3. PROPOSED SKA-RCRNNN-BASED SED MODEL

Fig. 1 illustrates the training procedure for the proposed SKA-RCRNN-based SED model, in which a two-stage mean-teacher model composed of SKA-RCRNN was trained by semi-supervised learning. A detailed explanation of the SKA-RCRNN architecture, SKA operation, and semi-supervising learning will be given in the following subsections.

3.1. Network architecture

Table 1 shows the proposed SKA-RCRNN architecture, which comprises one stem block, five residual convolutional (RC) blocks, and one recurrent neural network (RNN) block. Among the RC blocks, the first three blocks include frequency-wise and channel-wise selective kernel attention (fcwSKA), while the remaining two blocks include channel-wise selective kernel attention (cwSKA).

To begin with, all the input features of each audio clip are grouped to create a 625×128×1-dimensional spectral image, which is used as the input to the stem block. The stem block consists of two convolutional blocks with 16 and 32 kernels for the first and second convolutional blocks, respectively. Each convolutional block has 3×3 kernels with a stride of 1×1 and received batch normalization, gated linear unit (GLU) activation, and a 2×2 average pooling layer.

Next, the output of the stem block is processed by the first RC block that consists of a convolutional layer, fcwSKA, self-attention using a convolutional block attention module (CBAM) [6] and an average pooling layer, as shown in the table. Note here that the architecture and function of fcwSKA will be described in the next subsection. After that, the output of each RC block is passed to the next RC block, which resulted in a 156x1×128-dimensional feature map. Then, this feature map is applied to the RNN block, which consists of two bidirectional gated recurrent units (Bi-GRUs) to learn the temporal context information, where a rectified linear unit (ReLU) is used as an activation function for each Bi-GRU.

Finally, for the attainment of a strong label for each audio clip, the 156×256-dimensional output of the RNN block is processed by a fully connected (FC) layer and then by a sigmoid function, resulting in a 156×10-dimensional output, where 10 denotes the number of sound events to be detected. Note that the sound event category for each time and the timestamps for the detected events are obtained by thresholding the 156×10-dimensional output. For the attainment of a weak label for the consistency loss function used for the first stage of the proposed SED model, a weighted pooling layer is applied to the 156×10-dimensional output to obtain a 1×10-dimensional output.

3.2. Residual convolution with selective kernel attention

The selective kernel is composed of a split, fuse, and select step to account for various receptive fields by utilizing kernels with multiple sizes[5]. First, in the split step, feature maps are generated using the kernels that are different sizes. Next, the fuse step computes attention weights for different kernels via a sequence of processing that included squeezing, global average pooling, an FC layer,

Table 1: Network architecture of the proposed SKA-RCRNN-based SED model.

Name	Layers
Input layer	Input: log-mel spectrogram
Stem block	(3x3,Conv2D,@16,GLU,BN) 2x2 average pooling layer
	(3x3, Conv2D,@32,GLU,BN) 2x2 average pooling layer
Residual convolutional blocks	(3x3,Conv2D,@64,BN) fcw (fw(M,G,r=2,2,16)@64,ReLU,BN cw(M,G,r=2,8,16)@64,ReLU,BN) CBAM 1x2 average pooling layer
	((3x3,Conv2D,@128,BN) fcw (fw(M,G,r=2,2,16)@128,ReLU,BN cw(M,G,r=2,8,16)@128,ReLU,BN) CBAM 1x2 average pooling layer) x2
	((3x3,Conv2D,@128,BN) cw(M,G,r=2,8,16)@128,ReLU,BN CBAM 1x2 average pooling layer) x2
Recurrent neural network block	(128 BiGRU cells) x 2

and a softmax activation layer. Finally, the select step generates a merged feature map by a weighted sum of the feature maps from the different kernels by either channel-wise addition or frequency-wise addition.

Fig. 2 shows the RC block incorporating fwSKA, cwSKA, or both. Each fwSKA or cwSKA has 3×3 and 5×5 kernels with a stride of 1×1. As shown in the figure, CBAM-based attention [6] and a pooling layer are applied to the output of the feature map obtained by SKA.

3.3. Semi-supervised learning

The SKA-RCRNN-based SED model is trained on the basis of the two-stage training procedure using the mean-teacher and noisy-student models that were proposed in [8]. In other words, the proposed SKA-RCRNN model used the mean-teacher model and is trained to generate pseudo-labels for weakly labeled and unlabeled data. A Monte-Carlo (MC) dropout [9] technique from 10 samples with a 0.3 dropout probability is applied to obtain the pseudo-labels. Next, in the second training stage, the noisy-student model, which is also constructed by the SKA-RCRNN model, is trained using the pseudo-labels from the mean-teacher model of the first training stage, where a semi-supervised loss function defined in [8] is used. Data augmentation is also applied here using mix-up, SpecAugment, and time-frequency shifting.

4. EXPERIMENTAL RESULTS

4.1. Model training

The parameters of the SKA-RCRNN-based SED model were initialized in the first training stage using Xavier initialization, except for the biases, which were all initialized to zero. Then, the mini-batch-wise adaptive moment estimation optimization technique was utilized, in which dropout was also applied at a rate of 0.5. In addition, the learning rate was set according to the ramp-up strategy, with the maximum learning rate reaching 0.001 after 50 epochs.

The different versions of noisy-student models in the second training stage were trained according to different settings of the semi-supervised loss function parameter, β. In this work, β was set to 0.7, 0.9, or 1.0. In addition, five-fold cross-validations were applied for each β. Specifically, all of the data in the training dataset were divided into five folds, and four of them were used for training, while the remaining fold was used for validation. The learning rate was initially set to 0.001, and it was reduced by a simple learning rate schedule.

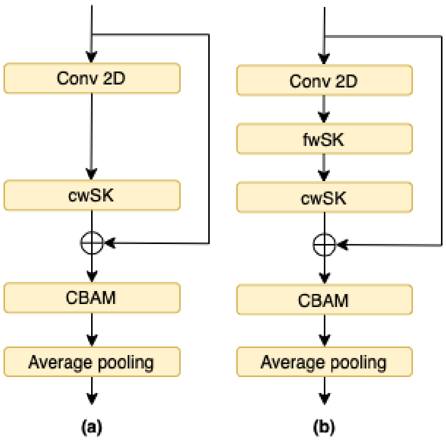

Figure 2: Residual convolution block with two different selective kernel attentions of (a) channel-wise selective kernel attention (cwSKA) and (b) frequency-wise and channel-wise selective kernel attention (fcwSKA).

4.2. Discussion

The performance of the proposed SED model was evaluated using the measures defined in the DCASE 2022 Challenge Task 4 [7], such as an event-based F1-score and a polyphonic sound detection score (PSDS) [10]. Table 2 compares the performance between the baseline[7], the EfficientNet-based model, and various versions of the proposed SED models on the validation dataset of the DCASE 2022 Challenge Task 4. As presented in Table 2, the pretraining, sampling, labeling and aggregation (PSLA)-based mean-teacher model scored 3.9%, 0.007, and 0.044 higher, whereas the SKA-RCRNN-based mean-teacher model scored 10.88%, 0.045, and 0.073 higher than the baseline's F1-score, PSDS-scenario 1, and PSDS-scenario 2, respectively. Furthermore, the noisy-student model (single model) scored 3.67%, 0.059, and 0.045 higher, in the F1-score, PSDS-scenario 1, and PSDS-scenario 2, respectively, than the SKA-RCRNN mean-teacher model. The Top1-10 ensemble models achieved further improvements in the noisy-student single model.

In this report, we used the EfficientNet-B2 from the PSLA single model that was trained using AudioSet for the audio tagging task [11]. Then, the EfficientNet-B2 was modified using an attention and aggregation module in the A^2-FPN [12]. Finally, the modified EfficientNet-B2 was fine-tuned using the DCASE 2022 Task 4 training dataset, which is referred to as EfficientNet-based mean-teacher model in the table.

As variants of the proposed SED model, the SKA-RCRNN-based SED model from the first training stage was evaluated as shown in the third row of the table, and the proposed SKA-RCRNN-based mean-teacher model achieved a higher F1-score by 10.88% and 6.38% than the baseline and EfficientNet-based SED models, respectively. Moreover, the PSDS-scenario 1 and PSDS-scenario 2 scores of the proposed SED model were higher than those of the baseline and EfficientNet-based SED models.

Table 2: Comparison of performance metrics of the baseline and different versions of the proposed SED model on the validation dataset of the DCASE 2022 Challenge Task 4.

Model	Mean-teacher	Noisy-student	Event-based F1-score	PSDS-scenario 1	PSDS-scenario 2
Baseline: CRNN-based mean-teacher model[7] (Single model)	✓	-	42.90%	0.351	0.552
EfficientNet-based mean-teacher model (Single model)	✓	-	47.40%	0.360	0.596
SKA-RCRNN-based mean-teacher model (Single model)	✓	-	53.78%	0.396	0.625
SKA-RCRNN-based noisy student model (Single model)	✓	✓	57.45%	0.455	0.670
SKA-RCRNN-based noisy student model (Top1-10 ensemble)	✓	✓	58.22%	0.456	0.685

Next, the SKA-RCRNN-based SED model from the second training stage was evaluated, with the semi-supervised loss function parameter, β, being set to 0.7. As shown in the fourth row of the table, it provided a better F1-score and better PSDS scores that the SKA-RCRNN-based SED model from the first training stage. Finally, an SKA-RCRNN-based SED model was obtained by ensembling top 1-10 models from all the models by different cross-validations and different βs. As shown in the last row of the table, this ensemble model showed the best performance in F1-score and PSDS-scenarios 1 and 2 of the SKA-RCRNN-based models. Interestingly, improvements of the F1-score, PSDS-scenario 1, and PSDS-scenario 2 by 15.32%, 0.105, and 0.133, respectively, were achieved compared to the baseline given by the DCASE 2022 Task 4.

5. CONCLUSION

This report proposed an SKA-RCRNN-based SED model for the DCASE 2022 Challenge Task 4. The proposed SED model was trained by applying semi-supervised learning to a two-stage mean-teacher model architecture, in which the student and teacher models were constructed by SKA-RCRNN. In addition, a consistency loss and a semi-supervised loss were used for the mean-teacher model in the first and second stage, respectively. Moreover, an MC dropout was applied to the noisy-student model in the second stage for ensemble learning. The proposed SKA-RCRNN-based SED model was evaluated on the validation dataset of the DCASE 2022 Task 4. In addition, the performance of different ensemble models was investigated by making them from fivefold cross-validation SED models with various hyper-parameters. Consequently, the experiments showed that an ensemble model from the top 1-10 models in terms of F1-score improved the F1-score, PSDS-scenario 1, and PSDS-scenario 2 by 15.32%, 0.105, and 0.133, respectively, compared to the baseline given by the DCASE 2022 Task 4.

6. ACKNOWLEDGMENT

This work was supported in part by Hanhwa Techwin Co. Ltd., and 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).

7. REFERENCES

[1] S. Grollmisch, J. Abeßer, J. Liebetrau, and H. Lukashevich, “Sounding industry: Challenges and datasets for industrial sound analysis,” in *Proc. European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[2] Z. Zhao, S.-H. Zhang, Z.-Y. Xu, K. Bellisario, N.-H. Dai, H. Omrani, and B. C. Pijanowski, “Automated bird acoustic event detection and robust species classification,” *Ecological Informatics*, vol. 39, pp. 99–108, 2017.

[3] K. Drossos, S. Adavanne, and T. Virtanen, “Automated audio captioning with recurrent neural networks,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 374–378.

[4] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in *Proc. ICASSP*, 2017, pp. 776–780.

[5] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, “Scaper: A library for soundscape synthesis and augmentation,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 344–348.

[6] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proc. ECCV*, 2018, pp. 3–19.

[7] https://dcase.community/challenge2022/task-sound-event-detection-in-domestic-environments.

[8] 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.

[9] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *Proc. International Conference on Machine Learning*, 2016, pp. 1050–1059.

[10] Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, “A framework for the robust evaluation of sound event detection,” in *Proc. ICASSP*, 2020, pp. 61–65.

[11] https://github.com/YuanGongND/psla.

[12] R. Li, L. Wang, C. Zhang, C. Duan, and S. Zheng, “A2-FPN for semantic segmentation of fine-resolution remotely sensed images,” *International Journal of Remote Sensing*, vol. 43, no. 3, pp. 1131–1155, 2022.