# SEMI-SUPERVISED SOUND EVENT DETECTION SYSTEM FOR DCASE 2022 TASK 4

Technical Report

*Kexin He, Xin Shu, Shaoyong Jia, Yi He*

Bytedance AI Lab
{hekexin, shuxin.shu, jiashaoyong, heyi.hy}@bytedance.com

## ABSTRACT

In this report, we describe our submissions for the task 4 of Detection and Classification of Acoustic Scenes and Events (DCASE) 2022 Challenge: Sound Event Detection in Domestic Environments. Our methods are mainly based on two types of deep learning models: Convolutional Recurrent Neural Network with selective kernel convolution (SK-CRNN) and frequency dynamic convolution (FDY-CRNN). In order to prevent overfitting, we adopt data augmentation using mixup strategy, FilterAugment, Interpolation Consistency Training (ICT) and Shift Consistency Training (SCT). Besides, we utilize external data and pretrained model to further improve performance, and try an ensemble of multiple subsystems to enhance the generalization capability of our system. Our final systems achieve a PSDS1/PSDS2 score of 0.5331/0.8569 on development dataset.

***Index Terms—*** DCASE, sound event detection, mean teacher, semi-supervised learning

## 1. INTRODUCTION

In task 4 of DCASE 2022, our purpose is to predict not only the the event class but also the event time localization. This task is dedicated to sound events detection (SED). In this task, 10 domestic sound events are considered as the target events. The prediction results of detection are finally evaluated with poly-phonic sound event detection scores (PSDS) [1].

In this report, we propose the sound event detection system based on the offical baseline. The baseline utilizes the Convolutional Recurrent Neural Network (CRNN) as the model architecture and apply Mean Teacher (MT) [2]. In our proposed approach, there are four main improvements. Firstly, we apply several data augmentation operations such as ICT [3], SCT [4], FilterAugment [5] on both time and frequency axises of the spectrogram. Secondly, we utilize the SK-CRNN [6] and FDY-CRNN [7] as the model architecture instead of CRNN. These two models have adaptive kernels and provide more flexibility. Thirdly, we investigate the relationship between the labels of AudioSet [8] and the target 10 acoustic events. We add some strongly labeled and weakly labeled data from AudioSet using the mapping relationship. Fourthly, pretrained models play a role of embedding which is concatenated to the model. Audio Spectrogram Transformer (AST) [9] is a convolution-free, purely attention-based model while PANN [10] is a CNN based model. We make some improvements to the convolution structure in PANN and achieve better results. Two pretrained models both achieve good performance on the audio classification task. By the way, the latter two improvements are the main focus of the DCASE 2022 task 4.

## 2. METHODS

### 2.1. Data preprocessing

All audio are resampled to 16kHz and down sampled to mono. We use log-mel energies as acoustic feature and extract 128 dimensional log-mel spectrogram using 2048 STFT window with a hop length of 256. In order to deal with the variable lengths of audio, we set a maximum padding length. All shorter feature will be zero-padding to the padding length. When it is longer, it will be truncated. In this work, maximum padding length is set to 626.

### 2.2. Data augmentation

We apply frame shift, time mask, frequency mask, gaussian noise and mixup these five common operations in our system to increase the robustness. What is more, we also apply three other data augmentation methods called FilterAugment, ICT and SCT.

We apply the FilterAugment on the input spectrogram to mimic complex acoustic conditions. More implement details are available in [5]. ICT encourages the prediction at an interpolation of unlabeled data points to be consistent with the interpolation of the prediction at these data points. SCT encourages the prediction of time-shifted and frequency-shifted inputs to be consistent with time-shifted and frequency-shifted prediction. Thus, the loss function during training can be indicated in follow formula. The baseline loss contains binary cross-entropy (BCE) and mean squared error (MSE). In this system, ICT and SCT introduce additional loss function and they are also added into the total loss. $S_\theta$ and $T_{\theta'}$ denote student and teacher model, $d_i$ and $d_j$ denote data points, and $\theta$ is randomly sampled from a Beta distribution. $wf$, $sf$ and $st$ denote clip-level outputs with frequency shift, frame-level outputs with frequency shift, and frame-level outputs with time shift, respectively.

$$L_{baseline} = L_{w,BCE} + L_{s,BCE} + w(t)(L_{w,MSE} + L_{s,MSE})$$

$$w(t) = exp[-5(1 - \frac{t^2}{T})]$$

$$L_{ICT} = MSE(S_\theta(\lambda d_i + (1-\lambda)d_j), \lambda T_{\theta'}(d_i) + (1-\lambda)T_{\theta'}(d_j))$$

$$L_{SCT} = L_{wf,BCE} + L_{sf,BCE} + L_{st,BCE} + w(t)(L_{st,MSE})$$

$$Loss = L_{baseline} + L_{ICT} + L_{SCT}$$

### 2.3. Mean teacher

We utilize Mean-Teacher model [2] for semi-supervised learning. It is a combination of two models: a student model and a teacher model, having the same architecture. The student model is the one used at inference while the goal of the teacher is to help the student

Table 1: Description for submitted system

| system | external data | pretrained model | weak prediction | model count | PSDS1 | PSDS2 |
|---|---|---|---|---|---|---|
| 1 | | | | 8 | 0.4743 | 0.6917 |
| 2 | ✓ | ✓ | | 40 | 0.5206 | 0.7709 |
| 3 | ✓ | ✓ | | 16 | 0.5331 | 0.7625 |
| 4 | ✓ | ✓ | ✓ | 16 | 0.0714 | 0.8569 |

model during training. The teacher's weights are the exponential average of the student model's weights. More details are available in [11].

### 2.4. Neural network

The selective kernel (SK) network [12] is a dynamic selection mechanism in CNN that allows each neuron to adaptively adjust its receptive field size based on multiple scales of input information. We replace the regular convolution layer in the official baseline (CRNN model) with selective kernel unit. We call this model SK-CRNN. The architecture of SK-CRNN is similar to that in [6]. In order to improve physical inconsistency in regular convolution layer on SED task, [7] propose frequency dynamic convolution which applies kernel that adapts to frequency components of input. We also utilize this architecture (FDY-CRNN) as our SED model.

The CNN part is composed of 7 convolution layers with [32, 64, 128, 256, 256, 256, 256] filters. Each convolution layer is followed by batch normalization, ReLU, dropout and avg-pooling. The avg-pooling kernel is [[2, 2], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 2]]. And a bi-directional gated recurrent unit (Bi-GRU) is used to capture temporal context. Finally, two dense layers are applied to output prediction scores for each class. Moreover, this is a weakly labeled SED task, we need to aggregate the frame-level probabilities into a clip-level probability. Attention mechanism is widely used in SED models [13, 14, 15]. Some handcraft pooling function also achieve good performance [16, 17]. We use attention and linear function in final pooling layer.

### 2.5. External data

DCASE provides 3470 AudioSet strongly labeled clips with the target 10 events. We find out a mapping relationship between the AudioSet 527 categories and the target 10 categories. We statistic the mapping relationship for each time stamps labeling in 3470 clips and calculate the mapping rate for each 527 categories. First, we set 60% as the mapping rate threshold. Only mapping relationships with mapping rates larger than 60% can be filtered out. Second, we only filter out mapping relationships which have reasonable semantic relationships. There are 29 relationships left through these two guidelines. With these mapping relationships, we select the strongly labeled data in AudioSet which contains 29 original acoustic categories. As there are too many clips labeled with "Speech", we remove those whose label is only "Speech" to ensure the balance between acoustic events. There are external 5167 strongly labeled clips. They can act not only the role of strongly labeled data, but also the role of weakly labeled data.

### 2.6. Pretrained model

Convolutional neural network is popular in audio related tasks for spectrograms. PANN model [10] achieves the state-of-the-art performance (0.439 mAP on AudioSet) in CNN based architecture. We replace convolution in PANN with separated unidirectional convolution to improve performance and achieve 0.460 mAP on AudioSet. Then we add this model to our SED system and utilize a trainable RNN encoder to encode features of pretrained model to a fixed dim output. This output is regard as clip-level features and then concatenated with CNN features from SED model. Besides, AST [9] is the first convolution-free, purely attention-based model for audio classification. We also add the pretrained AST model to SED system in the same way.

## 3. EXPERIMENTS

### 3.1. Experiment setup

There are 1578 weakly labeled clips, 14412 unlabeled clips, 10000 synthetic strongly labeled clips and 8637 (3470+5167) real strongly labeled clips used in system development. And the input for our SED systems consists of the spectrogram feature and the embedding from pretrained model. Then, the SED system is trained with different kinds of data augmentation methods (including frame shift, time mask, frequency mask, mix-up, gaussian noise, FilterAugment, ICT and SCT) and model architectures (including SK-CRNN and FDY-CRNN). We train the whole system for 200 epochs and the learning rate warms up in the first 50 epochs with the initial learning rate of 0.001. The batch size is set to 64.

### 3.2. Evaluation metric

The primary metric is poly-phonic sound event detection scores [1]. This metric is based on the intersection between events. PSDS values are computed using 50 operating points (linearly distributed from 0.01 to 0.99). In order to test SED system for different scenarios, we set two different PSDS parameters. In scenario1, the system needs to react fast upon an event detection. The localization of the sound event is important. In scenario2, the system must avoid confusing between classes but the reaction time is less crucial than in the first scenario. More details are available in [18].

### 3.3. Model ensemble and submissions

The systems we submitted are shown in Table 1. The four systems adopt the model fusion strategy. System1 is trained without external data and pretrained model. System2 and system3 are ensembled average of 40 models and 16 models respectively. Compared with system3, system4 utilize weak prediction method [19] to obtain higher PSDS2 score. The best PSDS1 score is 0.4539, and the best PSDS2 score is 0.7879.

## 4. CONCLUSION

In this report, we present our methods used in the task 4 of DCASE 2022 Challenge. We adopt FilterAug, mixup, ICT and SCT for data augmentation. We apply two types of deep learning model including SK-CRNN and FDY-CRNN. Besides, we add external data and pretrained model to further improve performance. Our final systems achieve a PSDS1/PSDS2 score of 0.5331/0.8569 on development dataset.

## 5. REFERENCES

[1] Bilen Ç, Ferroni G, Tuveri F, et al, "A framework for the robust evaluation of sound event detection," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020: 61-65.

[2] Tarvainen A, Valpola H, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in Advances in neural information processing systems, 2017, 30.

[3] Verma V, Kawaguchi K, Lamb A, et al, "Interpolation consistency training for semi-supervised learning," arXiv preprint arXiv:1903.03825, 2019.

[4] Koh C Y, Chen Y S, Liu Y W, et al, "Sound event detection by consistency training and pseudo-labeling with feature-pyramid convolutional recurrent neural networks," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021: 376-380.

[5] Nam H, Kim S H, Park Y H, "Filteraugment: An acoustic environmental data augmentation method," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022: 4308-4312.

[6] Zheng X, Chen H, Song Y, "Zheng ustc teams submission for dcase2021 task4 semi-supervised sound event detection," DCASE2021 Challenge, Tech. Rep, 2021.

[7] Nam H, Kim S H, Ko B Y, et al, "Frequency Dynamic Convolution: Frequency-Adaptive Pattern Recognition for Sound Event Detection," arXiv preprint arXiv:2203.15296, 2022.

[8] Gemmeke J F, Ellis D P W, Freedman D, et al, "Audio set: An ontology and human-labeled dataset for audio events," in international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2017: 776-780.

[9] Gong Y, Chung Y A, Glass J, "Ast: Audio spectrogram transformer," arXiv preprint arXiv:2104.01778, 2021.

[10] Kong Q, Cao Y, Iqbal T, et al, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2880-2894.

[11] Turpault N, Serizel R, Salamon J, et al, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," 2019.

[12] Li X, Wang W, Hu X, et al, "Selective kernel networks," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 510-519.

[13] Kong Q, Xu Y, Wang W, et al, "Audio set classification with attention model: A probabilistic perspective," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018: 316-320.

[14] JiaKai L, "Mean teacher convolution system for dcase 2018 task 4," in Detection and Classification of Acoustic Scenes and Events, 2018.

[15] Shen Y H, He K X, Zhang W Q, "Learning how to listen: A temporal-frequential attention model for sound event detection," in International Speech Communication Association (INTERSPEECH), 2019: 2563-2567.

[16] Wang Y, Li J, Metze F, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019: 31-35.

[17] He K X, Shen Y H, Zhang W Q, "Hierarchical pooling structure for weakly labeled sound event detection," in International Speech Communication Association (INTERSPEECH), 2019: 3624-3628.

[18] `https://dcase.community/challenge2022`.

[19] Nam H, Ko B Y, Lee G T, et al, "Heavily augmented sound event detection utilizing weak predictions," arXiv preprint arXiv:2107.03649, 2021.