SEMI-SUPERVISED SOUND EVENT DETECTION SYSTEM FOR DCASE 2023 TASK 4

Technical Report

Yadong Guan, Qijie Shang,

Harbin Institute of Technology, School of Computer Science and Technology, Harbin, China
guanyadonghit@gmail.com, hit_sqj@stu.hit.edu.cn

ABSTRACT

In this report, we describe our submissions for the task 4 of Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 Challenge: Sound Event Detection in Domestic Environments. Our methods are mainly based on Convolutional Recurrent Neural Network. We propose to utilize sound activity detection (SAD) as an auxiliary task for sound event detection and use a multi-task learning approach to train the two tasks simultaneously, thus improving the model generalization performance. Moreover, we proposed a new local weak prediction to improve the PSDS2 index. To prevent overfitting, we adopt data augmentation using hard mixup, pitch shift, and time shift. Besides, we utilize external data and a pretrained model named Beats 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 0.523/0.890 on development dataset.

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

1. INTRODUCTION

Sound event detection (SED) is the task of recognizing sound events and locating them temporally in audio recordings. Due to the lack of data with frame-level annotations, semi-supervised sound event detection (SS-SED) [1, 2] has received extensive attention.

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) [3]. The baseline also provides pre-trained model named Beat to obtain embedding to help model training. In our proposed approach, there are two main improvements. Firstly, we adoped Sound Activity Detection (SAD) as an auxiliary task, and train SOOD and SED in a multi-task training manner. Since SAD is good at detecting event boundaries, SAD can help SED improve the performance of boundary detection through the multi-task training. The SAD labels can be automatically derived from SED labels. Similar to semi-supervised SED, we also train SAD in a semi-supervised way. Secondly, the existing weak prediction methods are beneficial to improve the PSDS2 index [4]. However, when the event duration in the audio is short, the label of the weakly predicted output indicates that the sound persists. This may result in a decrease in PSDS2 performance. We propose a new local weak prediction method to improve the PSDS2 metric. The proposed methods achieve good performance on the PSDS1 and PSDS2.

2. METHODS

2.1. Sound activity detection

Sound Activity Detection (SAD) is used to detect the presence of sound activity in audio. Both it and SED involve the detection of event boundaries. Therefore, the two tasks are very related. We use multi-task learning to train the two tasks at the same time, so as to improve the generalization ability of the model. During training, labels for supervised training of SAD can be automatically generated by SED. Moreover, we use weakly labeled and unlabeled data to train SAD in a semi-supervised training manner.

2.2. 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.

2.3. Mean teacher

We utilize Mean-Teacher model [3] 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 model during training. The teacher's weights are the exponential average of the student model's weights. More details are available in [1].

2.4. Neural network

The SED and SAD tasks in our method share common features extracted through a shared backbone. Once these shared features are obtained, they are fed into separate branches for SED and SAD, respectively. The backbone architecture consists of four convolutional layers with filter sizes of [32, 64, 128, 256]. Each convolutional layer is followed by batch normalization, Context gating, dropout, and average pooling. The average pooling kernels used are [[2, 2], [2, 2], [1, 2], [1, 2]]. The SED branch comprises three convolutional layers with filter sizes of [256, 256, 256]. The average pooling kernels for this branch are [[1, 2], [1, 2], [1, 2]]. On the other hand, the SAD branch consists of two convolutional layers with filter sizes of [256, 256]. The average pooling kernels used for this branch are [[1, 2], [1, 4]], To capture the temporal context, both branches employ a bi-directional Gated Recurrent Unit (Bi-GRU).

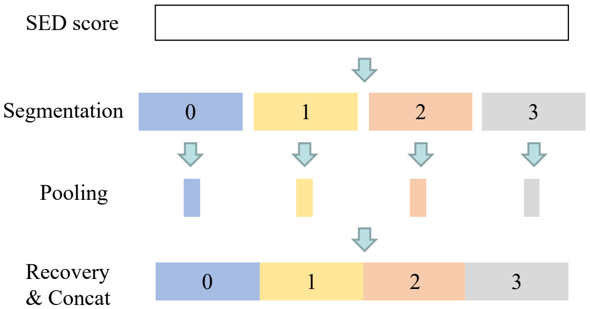

Figure 1: Local weak prediction.

Finally, two dense layers are applied to output prediction scores for SED and SAD. In addition, we also aggregate the frame-level SED scores into a clip-level score. We use attention pooling in final pooling layer. In addition, we also adopt a pre-trained model beat [5] to enhance model performance, and the specific settings are consistent with those in the baseline.

2.5. Local weak prediction

The existing weak prediction [4] regards the segment-level detection result as the result of each moment. The post-processing method effectively improved the PSDS2. However, when an sound event has a short duration, other frames where the event does not occur will also be assigned the class label, which may introduce additional false positives. To address this issue, we improve this approach and propose local weak predictions, which is shown in Fig. 1.

The specific steps are as follows. First, let $\boldsymbol{Y} \in \mathbb{R}^{T \times C}$ represent the frame-level posterior probability of the sound event, where T and C is the size of the time and event class dimension. Then, the maximum pooling operation is performed on the result, and the pooling kernel size and offset size are both 40. The obtained result dimension is $4 \times C$ Finally, we extend the result along the time dimension to bring it back to 156. In this way, local weak prediction results are obtained.

3. EXPERIMENTS

3.1. Experiment setup

There are 1578 weakly labeled clips, 14412 unlabeled clips, 10000 synthetic strongly labeled clips and 3470 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, hard mixup [6]). 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.

The primary metric is poly-phonic sound event detection scores [7]. This metric is based on the intersection between events. 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.

Table 1: Experimental results

extra data	pretrained model	weak prediction	model ensemble	PSDS1	PSDS2
				0.492	0.705
		√		0.105	0.839
√	√			0.517	0.782
√	√	√		0.113	0.885
√	√		√	0.523	0.790
√	√	√	√	0.115	0.890

3.2. Experimental results

The experimental results are shown in Table 1. Without additional data and pre-trained models, the PSDS1 and PSDS2 of the model are 0.492 and 0.705, respectively. After adding the local weak prediction, the PSDS2 index increased to 0.839. After adding additional data and pre-trained models, the PSDS1 and PSDS2 of the model are 0.517 and 0.782, respectively. After model integration, the best PSDS1 and PSDS2 performances are 0.523, 0.890.

4. CONCLUSION

In this report, we present our methods used in the task 4 of DCASE 2023 Challenge. We employ multi-task learning to simultaneously train SED and SAD, and propose local weak predictions Besides, we add external data and pretrained model to further improve performance. Our final systems achieve a PSDS1/PSDS2 score of 0.523/0.890 on development dataset.

5. REFERENCES

[1] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[2] K. He, X. Shu, S. Jia, and Y. He, "Semi-supervised sound event detection system for dcase 2022 task 4."

[3] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NIPS*, 2017, pp. 1195–1204.

[4] H. Nam, B. Ko, G. Lee, S. Kim, W. Jung, S. Choi, and Y. Park, "Heavily augmented sound event detection utilizing weak predictions," *arXiv preprint arXiv:2107.03649*, 2021.

[5] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," 2022.

[6] N. Shao, E. Loweimi, and X. Li, "RCT: Random consistency training for semi-supervised sound event detection," in *Proc. Interspeech*, 2022, pp. 1541–1545.

[7] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Threshold independent evaluation of sound event detection scores," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 1021–1025.