LI_USTC TEAM'S SUBMISSION FOR DCASE 2023 CHALLENGE TASK4A

Technical Report

Kang Li, Pengfei Cai, Yan Song

National Engineering Research Center of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China.
likang0311@mail.ustc.edu.cn, songy@ustc.edu.cn

ABSTRACT

In this technical report, we present our submissions for DCASE 2023 challenge task4a. We mainly study how to fine-tune patchout fast spectrogram transformer (PaSST) for sound event detection task (PaSST-SED). Firstly, we fine-tune PaSST with weakly-labeled DESED dataset. Task-aware fine-tuning (TAFT) and self-distillated mean teacher (SdMT) are used as fine-tuning strategies, TAFT helps exploit both local and semantic information from PaSST and SdMT helps train a robust model with soft knowledge distillation. Secondly, we fine-tune PaSST with pseudo-labeled DESED with pseudo labels from DCASE2022 rank1, mix-up is used to mix the audios with true or pseudo labels. Besides, when test with PaSST-SED model, slide window clipping (SWC) is used to compensate the temporal resolution loss of PaSST feature. We also evaluate post-processing methods including median-filtering and max-filtering. Experiments on the DCASE2023 task4a validation dataset demonstrate the effectiveness of the techniques used in our systems. Specifically, our systems achieve the best PSDS1/PSDS2 of 0.5624/0.8990.

1. INTRODUCTION

Sound event detection (SED) is the task to detect both the onset and offset of a sound event and classify its categories. It has wide applications for real-world systems including smart home devices [1], and automatic surveillance [2]. Since DCASE2018, due to the difficulty of manually annotating sound events, only a small quantity of weakly-labeled data is available, to utilize large-scale unlabeled data, semi-supervised learning (SSL) based SED methods have been explored in the past. Mean teacher (MT) [3] has built a strong SSL baseline, and other SSL methods such as interpolation consistency training (ICT) [4], shift consistency training (SCT) [5], and confident mean teacher (CMT) [6] have been proposed to exploit unlabeled data efficiently. From DCASE2019 to DCASE2021 [7, 8], synthetic data with accurate time-stamps have been proposed and get larger and larger, some methods untilizing the strongly-labeled data achieved state-of-the-art performance [9, 10, 11]. Considering the domain gap between synthetic and real audio data, [12, 13] explore the domain adaptation methods to exploit synthetic strong-labeled data efficiently.

In DCASE2022, several researches on exploiting external large-scale weakly-labeled AudioSet [14] data have greatly improved the detection performance of SED systems. For example, the forward-backward CRNN (FB-CRNN) and Bi-directional CRNN (Bi-CRNN) [15] are firstly pretrained on AudioSet, then they are fine-tuned in a self-training manner, which achieves the first rank in DCASE2022 task4. Xiao [16] study how to fine-tune pretrained AT models such as audio neural network (PANN) [17] and audio spectrogram transformer (AST) [18]. In our previous work AST-SED [19], the frequency-wise transformer encoder (FTE) and local GRU decoder (LGD) are proposed to effectively fine-tune AST for SED, it helps to extract a better temporal sequence, and produces a high-temporal-resolution representation, which is beneficial for SED task. AST-SED shows that pretrained AST model can be well transferred to SED task with no need to redesign or retrain the AST model.

In this year's challenge (i.e., DCASE2023), the main research is also how to exploit large-scale external data. We follow our previous work [19], and further study how to transfer Patchout faSt Spectrogram Transformer (PaSST) [20] model to sound event detection (SED) task. There are two main points to our work, firstly we study how to fine-tune PaSST with weakly-labeled DESED [7] dataset, and we apply the task-aware fine-tuning (TAFT) and self-distillated mean teacher (SdMT) to exploit the pretrained PaSST adequately. Secondly, we study how to fine-tune PaSST with pseudo-labeled DESED with pseudo labels from [15], we mix the audios with true or pseudo label to make the model not overfit to the data with noisy pseudo labels.

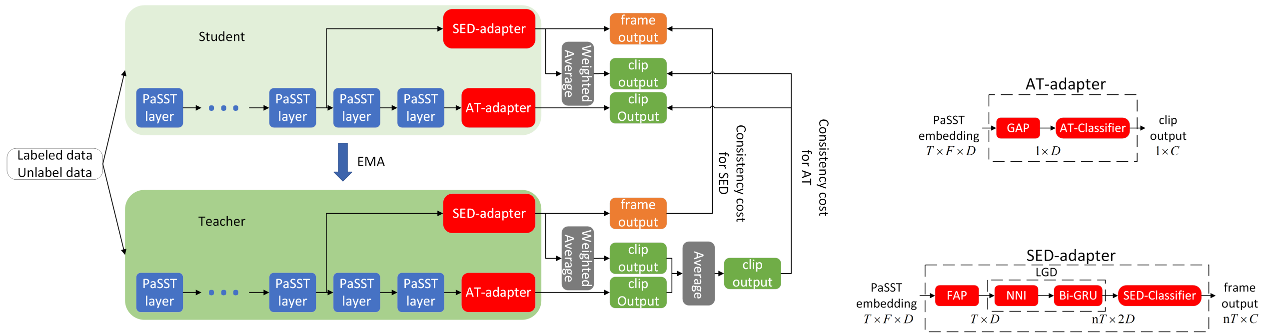
2. METHODS

2.1. Fine-tune PaSST with weakly-labeled DESED

2.1.1. Task-aware fine-tuning

As shown in Figure 1(a), in the task-aware fine-tuning (TAFT), given the output of PaSST, we use two task-adapters including SED-adapter and AT-adapter to transfer PaSST for SED or AT task respectively. As shown in Fiture 1(b), the SED-adapter consists of: (1) frequency-wise average pooling (FAP) to extract a frame-level representation, (2) local GRU decoder (LGD) [19] to produce a high-temporal-resolution representation, (3) SED classifier to produce frame-level SED output. The AT-adapter consists of: (1) Global average pooling (GAP) to extract a clip-level representation, (2) AT classifier to produce a clip-level output. The SED-adatper is attached to shallower layer to exploit local information while AT-adatper is attached to deeper layer to exploit semantic information. The AT-adatpers helps produce more accurate clip-level prediction to guide the SED-adapter learning. The loss function of SED-adatper is defined as follows,

$$L_{SED} = L^{sed}_{BCE,frame} + \lambda^{sed}_1 L^{sed}_{BCE,clip} + \lambda^{sed}_2 L^{sed}_{MSE,frame} + \lambda^{sed}_3 L^{sed}_{MSE,clip} \quad (1)$$



(a) Task-aware fine-tuning (b) Task-aware adapter

Figure 1: Task-aware fine-tuning.

where $L_{BCE,frame}^{sed}$ denotes frame-level classification BCE loss for strongly-labeled data, $L_{BCE,clip}^{sed}$ denotes clip-level classification BCE loss for weakly-labeled data, $L_{MSE,frame}^{sed}$ and $L_{MSE,clip}^{sed}$ denote frame-level and clip-level teacher-student consistency MSE loss for unlabeled data respectively. The weight λ_1^{sed}, λ_2^{sed}, λ_3^{sed} is set to 0.5, 2, 2 respectively. The clip-level output y_{clip} is a weighted average from frame-level output y_{frame} with linear-softmax pooling [21],

$$y_{clip} = \sum_{i=0}^{T} y_{frame,i}^2 / \sum_{i=0}^{T} y_{frame,i} \tag{2}$$

where i denotes th i^{th} frame. The loss fuction of AT-adatper is defined as follows,

$$L_{AT} = L_{BCE,clip}^{at} + \lambda_1^{at} L_{MSE,clip}^{at} \tag{3}$$

where the weight λ_1^{at} is set to 1, $L_{BCE,clip}^{at}$ denotes classification BCE loss for weakly-labeled data and $L_{MSE,clip}^{at}$ denotes clip-level teacher-student consistency MSE loss for unlabeled data. Total loss is as follows,

$$L_{task-aware} = L_{SED} + \lambda_{AT} L_{AT} \tag{4}$$

where L_{SED} and L_{AT} are same as Eqn. (1) and Eqn. (3) respectively, λ_{AT} is set to 2.

2.1.2. *Self-distillated mean teacher*

As the timestamps are hard to determined, the strong-label in the training data may be noisy, soft label from teacher may contain more information and deserved to be explored further, we propose self-distillated mean teacher (SdMT) to train a robust vice-student model with knowledge distillation (KD). As shown in Figure 2, same as mean teacher, the main-student is trained with labeled data, the teacher model is an EMA from main-student model, the teacher model guide the main-student learning with consistency regularization for unlabeled data, we introduce a vice-student, and distill the knowledge from teacher to vice-student with soft KD. The KD loss is as follows,

$$\begin{aligned} L_{kd,soft} = & MSE(\delta(z_{s,frame}), \delta(z_{t,frame}/\tau)) \\ & + \lambda_{clip} MSE(\delta(z_{s,clip}), \delta(z_{t,clip}/\tau)) \end{aligned} \tag{5}$$

where $z_{s,frame}$, $z_{t,frame}$, $z_{s,clip}$, $z_{t,clip}$ denotes student frame-level logits, teacher frame-level logits, student clip-level logits, teacher clip-level logits respectively, δ denotes sigmoid activation function, and the temperature τ is set to 1.

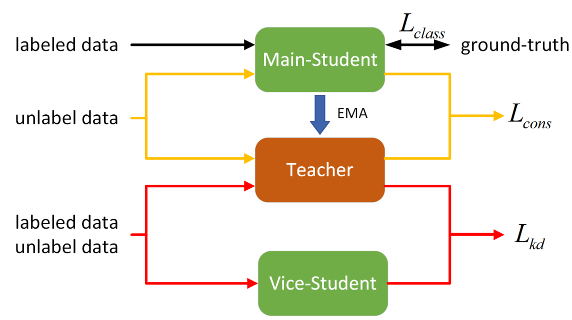


Figure 2: Self-distillated mean teacher (SdMT).

2.2. Fine-tune PaSST with pseudo-labeled DESED

When training with pseudo-labeled DESED, we do not apply mean teacher, and the model is trained in a supervised learning manner. However, as the pseudo label (PL) is noisy, we apply mix-up [22] to mix the audios in a training batch which contains true or pseudo labels, which may reduce the overfitting to noisy labels. The loss fuction is as follows,

$$L_{PL} = L_{BCE,frame} + \lambda_{clip} * L_{BCE,clip} \tag{6}$$

where $L_{BCE,frame}$, $L_{BCE,clip}$ denotes frame-level and clip-level BCE loss respectively. λ_{clip} is set to 0.5.

2.3. Post processing

2.3.1. *Median filtering and Max filtering*

Median filtering (MedianF) has been explored in the past challenges, window size are tuned individually for each event class to achieving the best event-based F1-score [15, 23]. Median filtering helps achieve better frame-level detection performance. We also propose to use Max filtering (MaxF) to achieve better segment-level detection performance. Specifically, we enlarge the window size with a ratio of 10, then we apply max filtering.

2.3.2. *Slide window clipping*

While the LGD block helps produce high-temporal resolution features in the training phase, we let the PaSST model producing high temporal resolution features itself in the test phase which may help

LGD produce better representations, specifically, when test, the input spectrogram is clipped to many sub-spectrograms along temporal axis with a window size and stride, they are feed to PaSST and further aggregated after NNI, the GRU helps decode better representations with the slide window clipping (SWC) operations. The window size is 516 and stride is 32.

3. EXPERIMENTS SETUP

3.1. Dataset

Experiments is conducted on DCASE2023 task4 development set (DESED) [7]. The training dataset contains: 1578 weakly-labeled clips, 3470 strongly-labeled clips, 10000 synthetic-strongly-labeled clips, and 14412 unlabeled in-domain clips. The validation dataset consists of 1168 strongly-labeled clips.

3.2. Feature Extraction

A 32kHz audio input waveform is first converted into 128-dimensional log Mel spectrogram features with a window size of 25ms and frame shift of 10ms. As a result, each 10-second sound clip is transformed into a 2D time-frequency representation with a size of (1000×128), then it shares same normalization as [24]. Frequency mask [25] and Mix-up are used for data augmentation.

3.3. Experimental Settiongs

The model is trained over 20 epochs with the AdamW [26] optimizer, and a ratio of 1:1:2:2 for strong, synthetic-strong, weak and unlabeled data is used for each batch. Learning rates (lr) are set to 5e-6, 1e-4 for pre-trained PaSST and the task-aware adapters. During training, the lr is constant for the first 10 epochs, then reduced with exponential-down schedule to 5e-7,1e-5 for the last 10 epochs. When using SdMT, the main-student and teacher are firstly trained over 20 epochs, then the vice-student is trained over another 20 epochs with the aforementioned settings. True Polyphonic Sound detection Score (PSDS) [27] is used to evaluate fine-grained SED performance, the scenario1 (PSDS1) is used to evaluate the fine-grained performance while scenario2 (PSDS2) is used to evaluate the coarse-grained performance.

4. RESULTS

- Ensemble1: an ensemble of 6 single1 models (our submitted system1).
- Ensemble2: an ensemble of 4 single2 models (our submitted system2).
- Ensemble3: an ensemble of 4 single models, trained with TAFT, Asymetrical focal loss (AFL), Mixup, SW and MedianF (our submitted system3).
- Ensemble4: replacing the MeidanF with MaxF in Ensemble3 (our submitted system4).
- Single1: we term this model as TAFT+SdMT+SWC+MedianF which denotes the task-aware fine-tuning (TAFT), self-distillated mean-teacher (SdMT), Slide window clipping (SW) and median filtering are used (our submitted system5) .
- Single2: we term this model as PL+Mixup+SWC+MedianF which denotes the pseudo labeling (PL), Mixup, Slide window clipping (SWC) and median filtering are used (our submitted system6).
- Single3: we term this model as SKCRNN, where the model structure is same as our DCASE2021 submission [28] and no external data is used to train this model. Training settings are same as [29] (our submitted system7).

Table 1: Submitted systems' performances on validation set.

Model	PSDS1	PSDS2
Baseline	0.5000	0.7620
Single1	0.5550	0.7914
Single2	0.5524	0.7947
Single3	0.4512	0.6622
Ensemble1	0.5624	0.7953
Ensemble2	0.5542	0.7990
Ensemble3	0.5585	0.7984
Ensemble4	0.0930	0.8990

As shown in Table 1, With TAFT and SdMT, the model (single1) achieves 0.5550 PSDS1 and 0.7914 PSDS2. With pseudo labels, the model (single2) achieves competitive results of 0.5524 PSDS1 and 0.7947 PSDS2. Ensemble model achieves higher results. After using MaxFiltering (Ensemble4), our model achieves the best PSDS2 of 0.8990 which shows the PSDS2 reflect the segment-level performance.

5. REFERENCES

[1] A. Southern, F. Stevens, and D. Murphy, "Sounding out smart cities: Auralization and soundscape monitoring for environmental sound design," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3880–3880, 2017.

[2] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.* IEEE, 2005, pp. 158–161.

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

[4] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, A. Solin, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," *Neural Networks*, vol. 145, pp. 90–106, 2022.

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

[6] S. Xiao, "Multi-dimensional frequency dynamic convolution with confident mean teacher for sound event detection," *arXiv preprint arXiv:2302.09256*, 2023.

[7] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," 2019.

[8] N. Turpault, S. Wisdom, H. Erdogan, J. Hershey, R. Serizel, E. Fonseca, P. Seetharaman, and J. Salamon, "Improving sound event detection in domestic environments using sound separation," *arXiv preprint arXiv:2007.03932*, 2020.

[9] Z. Shi, L. Liu, H. Lin, R. Liu, and A. Shi, "Hodgepodge: Sound event detection based on ensemble of semi-supervised learning methods," *arXiv preprint arXiv:1907.07398*, 2019.

[10] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning convolution system for dcase 2019 task 4," *arXiv preprint arXiv:1909.06178*, 2019.

[11] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution augmented transformer for semi-supervised sound event detection," in *Proc. Workshop Detection Classification Acoust. Scenes Events (DCASE)*, 2020, pp. 100–104.

[12] L. Yang, J. Hao, Z. Hou, and W. Peng, "Two-stage domain adaptation for sound event detection," in *Proc. Detection Classification Acoust. Scenes Events Workshop*, 2020, pp. 41–45.

[13] X. Zheng, Y. Song, L.-R. Dai, I. McLoughlin, and L. Liu, "An effective mutual mean teaching based domain adaptation method for sound event detection," *Proc. Interspeech 2021*, pp. 556–560, 2021.

[14] 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[15] J. Ebbers and R. Haeb-Umbach, "Pre-training and self-training for sound event detection in domestic environments," DCASE, Tech. Rep., June 2022.

[16] S. Xiao, "Pretrained models in sound event detection for dcase 2022 challenge task4," DCASE, Tech. Rep., June 2022.

[17] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 28, pp. 2880–2894, 2020.

[18] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.

[19] K. Li, Y. Song, L.-R. Dai, I. McLoughlin, X. Fang, and L. Liu, "Ast-sed: An effective sound event detection method based on audio spectrogram transformer," *arXiv preprint arXiv:2303.03689*, 2023.

[20] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," *arXiv preprint arXiv:2110.05069*, 2021.

[21] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *IEEE ICASSP*, 2019, pp. 31–35.

[22] Y. N. D. D. L.-P. Hongyi Zhang, Moustapha Cisse, "mixup: Beyond empirical risk minimization," *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[23] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution-augmented transformer for semi-supervised sound event detection," DCASE, Tech. Rep., June 2020.

[24] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient Training of Audio Transformers with Patchout," in *Proc. Interspeech 2022*, 2022, pp. 2753–2757.

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[27] 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.

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

[29] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," *arXiv preprint arXiv:2203.15296*, 2022.