A TWO-STAGE TRAINING METHOD FOR DCASE 2022 CHALLENGE TASK4

Technical Report

Kang Li, Xu Zheng, Yan Song

National Engineering Research Center of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China.
{likang0311, zx980216}@mail.ustc.edu.cn, songy@ustc.edu.cn

ABSTRACT

The goal of DCASE 2022 CHALLENGE TASK4 is to evaluate systems for the detection of sound events using real data either weakly labeled or unlabeled, simulated data that is strongly labeled and external data. In this technical report, we present a two-stage learning strategy based method to explore synthetic strong data and real strong data (from AudioSet). Specifically, a CRNN model is used as the baseline SED system for this year's challenge. According to different supervisory signals from weakly-labeled and strongly-labeled data, the frame-level and clip-level tasks (*i.e.* SED and Audio Tagging (AT)) are designed. In the first stage, the model is trained on weakly labeled, unlabeled and synthetic data with strong labels under the semi-supervised learning framework, *i.e.* Mean Teacher (MT). There are two types of MT, including frame-level MT and clip-level MT, corresponding to the subsets with different supervisory signals. In the second stage, a new model is trained using pseudo-labeling scheme, in which the pre-trained teacher model is utilized to provide the pseudo-label of the real weakly and unlabeled data. Furthermore, we explore the strongly labeled real data as external one in both stages. Results on the DCASE2022 Task4 validation set verify the effectiveness of our proposed method with PSDS1 and PSDS2 of 0.479 and 0.785, outperforming the baseline results of 0.351 and 0.552 respectively.

***Index Terms*—** Sound Event detection, two-stage, mean teacher, pseudo labeling, strongly labeled real data

1. INTRODUCTION

Sound event detection(SED) aims 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]. Due to the difficulty of manually annotating sound events, only weakly-labeled and unlabeled dataset are available in DCASE2018 [3]. Semi-supervised learning (SSL) methods such as mean teacher [4] are introduced to SED and achieved relatively good results, but strongly-labeled data are still in urgent need as the field evolves. Recently, synthetic data with accurate time-stamps have been proposed and get larger and larger from DCASE2019 to DCASE2021 [5, 6], some methods untilizing the strongly-labeled data achieved state-of-the-art performance [7, 8, 9]. However, these methods ignore the domain gap between synthetic and real audio data. Although several domain adaptation methods [10, 11] have been proposed for dealing with this problem, they just achieved small improvement. In DCASE2022, external data such as AudioSet [12] are allowed to train SED model, which increase the potential for improving performance. Thanks DACSE organizers for collecting the in-domain sub-dataset from AudioSet Strong [13], strongly-labeled real audio join in the field for the first time. From now on, four different types of data (weakly-labeled, unlabeled, strongly-labeled, synthetic) are available and need to be further studied. In this year's challenge, we aim to evaluate strongly-labeled and synthetic data with semi-supervised or supervised training. We propose a two-stage learning method consists of mean teacher and pseudo labeling to achieve our goal.

2. PROPOSED METHOD

Fig 1 shows our proposed two-stage training pipeline with CRNN as backbone and two training strategies include mean teacher[4] and pseudo labeling[14]. Details will be given in the following subsections.

2.1. First stage: semi-supervised training with mean teacher

Considering the missing labels of weakly-labeled and unlabeled data, semi-supervised training methods are used to train SED models and mean teacher (MT) is the most popular one. Under mean teacher structure, a input sample goes through two branches named student and teacher. The student model is trained with available true labels as well as pseudo labels from a teacher model, while the teacher model is updated by Exponential moving average (EMA). The input to student model is perturbed with frequency masking to regularize the model. In this stage, we have three goals:

(1): evaluate CRNN models trained with different dataset combinations under mean teacher structure.

(2): On the basis of (1), explore the effect of adding other two strategies: Selective Kernel (SK) unit [15] and Regularize GRU (R-GRU). R-GRU will be present at the end of this section.

(3): obtain the optimal model from the exploration of (1) and (2) as the teacher model for the next stage.

Total training loss L_{1stage} is defined as:

$$L_{1stage} = L_{class,BCE} + r(t)L_{MT,MSE} + \lambda L_{R-GRU,MSE} \quad (1)$$

where r(t) is the mean teacher MSE loss weight which first ramp up to 2 and then keep it. λ is a fixed R-GRU loss weight and we set it to 15 if use R-GRU.

2.2. Second stage: supervised training with pseudo labeling

The second stage contains two sub-stages. Firstly, weakly-labeled and unlabeled data get frame-level labels from the teacher model obtained in the first stage. Secondly, a new CRNN model are

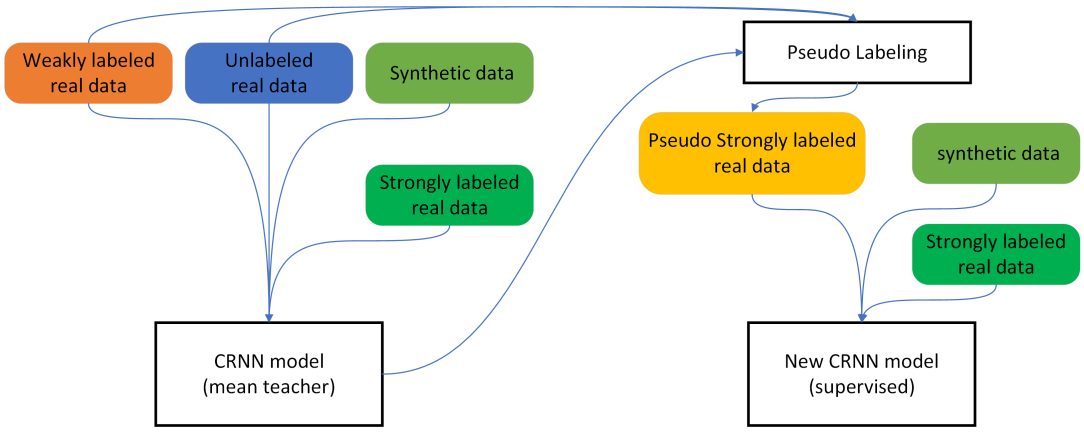

Figure 1: two-stage training pipeline.

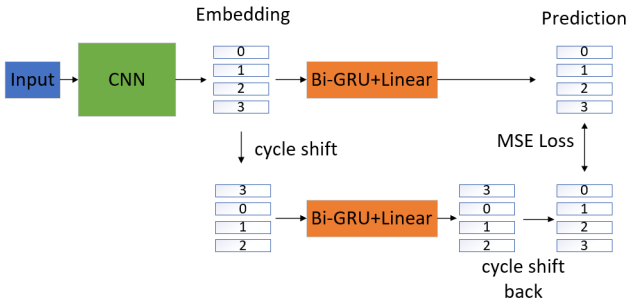

Figure 2: regularize GRU

trained with supervion. SK unit, frequency masking and R-GRU are still used in this stage. Our main goal is to evaluate CRNN models trained with pseudo-labeled and either synthetic or real strong dataset.

Total training loss L_{2stage} is defined as:

$$L_{2stage} = L_{class,BCE} + \lambda L_{R-GRU,MSE} \quad (2)$$

2.3. Regularize GRU.

It is well known that the position of a sound event within the clip is a large impact on detection performance [16], we think the overfitting GRU in CRNN model should be responsible for this phenomenon. We propose a regularization method to GRU: the output generated from different sequence modeling process should be same. Inspired by time-shifting and SCT [17], we apply cycle shift to change sequence modeling process as Fig 2 shows.

3. EXPERIMENTS

3.1. Dataset and Feature Extraction

In this year's challenge, we use DESED train dataset [5] and a strongly labeled real dataset from AudioSet [13]. DESED train data consists of three parts: 1) weakly labeled dataset (1578 clips), 2) unlabeled in-domain dataset (14412 clips), 3) strongly labeled synthetic dataset (10000 clips). The strongly labeled real dataset (3780 clips) act as external data for training and exploration. The input features used in the proposed system are log-mel spectrograms, which are extracted from the audio signal resampled to 16000 Hz. The log-mel spectrogram uses 2048 STFT windows with a hop size of 313 and 128 Mel-scale filters. As a result, each 10-second sound clip is transformed into a 2D time-frequency representation with a size of (512×128).

3.2. Experimental Settings

In both stages, the neural networks are trained using the Adam optimizer [18], with a maximum learning rate of 0.001. Total epochs are 100 and the learning rate ramp up during the first 20 epochs and ramp down during the remaining epochs. Batchsize is set to 64.

4. RESULTS AND ANALYSIS

4.1. Results of the first stage

4.1.1. evaluation for different dataset combination

we first evaluate several CRNN models trained with different dataset combinations. We also conduct some experiments with only weakly-labeled and strongly-labeled data for analysis (without mean teacher). For faster computation, we do not use SK unit and R-GRU in these experiments. We choose event-based F1 [19] and PSDS [20] as main metrics. Dataset combinations and results are shown in Table 1.

With supervised training, W+SR gets worse results among all metrics than W+SS with event-based F1, PSDS1, PSDS2 decreased by 0.01, 0.0498, 0.021 respectively. Noisy labels and smaller data volumn may be two reasons why SR is worse than SS. W+SS+SR is the best dataset combination. With mean teacher, the model trained with W+SS+SR gets the highest results. It is worth noting that W+U+SS+SR gets worse results than W+U+SR which may be a opposite conclusion compared with supervised training. One possible explanation is that there is enough samples in W+U+SR and further adding SS introduced domain gap problems.

Table 1: evaluation for different dataset combinations within Baseline. Baseline is a traditional CRNN model whose CNN part is stacked VGG blocks and RNN part is Bi-GRU.

Training Method	Data	EB-F1	PSDS1	PSDS2
Supervised	W[1]+SS[3]	0.4598	0.3646	0.6621
	W+SR[4]	0.4496	0.3148	0.6411
	W+SS+SR	0.4977	0.3745	0.7096
Mean Teacher	W+U[2]+SS	0.4840	0.3897	0.6903
	W+U+SR	0.5115	0.4056	0.7009
	W+U+SS+SR	0.4883	0.3883	0.7007

[1] W: weakly labeled real dataset.
[2] U: unlabeled real dataset.
[3] SS: strongly labeled synthetic dataset.
[4] SR: strongly labeled real dataset.

Table 2: evaluation for two strongly labeled dataset.

Data	Model	EB-F1	PSDS1	PSDS2
W+U+SS	Baseline	0.4840	0.3897	0.6903
	+SK	0.5191	0.4162	0.6945
	+SK+R-GRU	0.5387	0.4313	0.7028
W+U+SR	Baseline	0.5115	0.4056	0.7009
	+SK	0.5160	0.4083	0.6994
	+SK+R-GRU	0.5332	0.4286	0.7134

4.1.2. Evaluation for two strongly labeled data with SK unit and R-GRU

We further evaluate strongly labeled synthetic data and strongly label real data after applying SK unit and R-GRU. Results are shown in Table 2. After using SK and R-GRU, W+U+SS achieved 0.5387 EB-F1, 0.4313 PSDS1 and 0.7028 PSDS2, W+U+SR achieved 0.5332 EB-F1, 0.4286 PSDS1 and 0.7134 PSDS2. It is interesting that SR is no longer better than SS, which is a different conclusion from the baseline (SR achieved higher results among all metrics than SS within baseline) . We believe that the noisy labels in SR limits its potential. Therefore, we make a deeper digging to campare SS and SR with pseudo-labeled W+U in the next stage. The CRNN model with SK unit, which trained with W+U+SS and R-GRU, is choosen as teacher model for the next stage.

4.2. Results of the second stage

Firstly, W+U gets frame-level pseudo labels from the teacher model described above, then, W+U together with SS or SR are used to train a new model. When test, we ensemble several single models from different random seeds to get a higher results. All results are shown in Table 3.

When using SS, the second stage model get 0.4446 PSDS1 and 0.7331 PSDS2, the ensemble model get 0.4554 PSDS1 with T=3 (T is temperature as we used last year [15]) and 0.7778 PSDS2 with T=15. When using SR, the second stage model get 0.4582 PSDS1 and 0.7412 PSDS2, the ensemble model get 0.4790 PSDS1 with T=3 and 0.7852 PSDS2 with T=15. Both the single model and ensemble model trained with SR get better results than the model trained with SS. Data distribution of SS is not same as real data, so the second stage model trained with all real data(W+U+SR) can get higher result.

Table 3: evaluation for two-stage learning strategy. SK: Selective Kernel. R-GRU: Regularize GRU. MT: mean teacher. PL: pseudo labeling. T: temperature applied on logits.

Data	Stage	Model	PSDS1	PSDS2
W+U+SS	1	SK+MT+R-GRU	0.4313	0.7028
	2	SK+PL+R-GRU	0.4446	0.7331
	Test	*Ensemble_3model (T=3)	0.4554	0.7260
	Test	*Ensemble_3model (T=15)	0.4358	0.7778
W+U+SR	2	SK+PL+R-GRU	0.4582	0.7412
	Test	*Ensemble_10model (T=3)	0.4790	0.7352
	Test	*Ensemble_10model (T=15)	0.4619	0.7852

5. SUBMISSION SYSTEM

Models with * in Tabel 3 are our four submissions on DCASE 2022 Task4, including two models without external data and two models with external data.

6. 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] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," *arXiv preprint arXiv:1807.10501*, 2018.

[4] 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.

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] S. Hershey, D. P. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 366–370.

[14] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896.

[15] 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.

[16] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 86–90.

[17] 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.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[20] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 61–65.