DCASE 2022 TASK4 CHALLENGE TECHNICAL REPORT

Technical Report

Junyong Hao, Shunzhou Ye, Cheng Lu, Fei Dong, Jingang Liu
UNISOC, Chongqing, China.
junyong.hao@unisoc.com, jingang.liu@unisoc.com

ABSTRACT

This report proposes a polyphonic sound event detection (SED) method for the DCASE 2022 Challenge Task 4-Sound Event Detection in Domestic Environments. We use the dataset of DESED to train our model, contains strongly labeled synthetic data, large unlabeled data, weakly labeled data and strongly labeled real data. To perform this task, we propose a DACRNN network for joint learning of SED and domain adaptation (DA).We consider the impact of the distribution within a single sound on the generalization performance of the model by mitigating the impact of complex background noise on event detection and the self-correlation consistency regularization of clip-level sound event classification, these make the intra-domain of a single sound smoother; for cross-domain adaptation, adversarial learning through feature extraction network with weighted frame-level domain discriminator. Experiments on the DCASE 2022 task4 validation dataset and public-evaluation dataset demonstrate the effectiveness of the techniques used in our system. Specifically, PSDS1 scores of 0.448 and PSDS2 scores of 0.853 are achieved for validation dataset, PSDS1 scores of 0.553 and PSDS2 scores of 0.836 are achieved for public-evaluation dataset.

Index Terms— Domain adaptation, Sound event detection, Adversarial learning, Semi-supervised learning

1. INTRODUCTION

Compared with tedious sound event accurate labeling in real data, it is much easier to collect a certain number of sound event samples and background sound. Synthesizing these sound event samples and background sound to generate high-quality labeled audio sequences for supervised SED is sensible.

In the Challenge of Detection and Classification of Acoustic Scenes and Events (DCASE), supervised SED methods were tested both on synthetic and real-life audio datasets. The results demonstrated that sound event detection in a realistic setting was difficult. Undoubtedly, SED models trained using synthetic sequences perform robust less under real scenarios due to the statistical distribution mismatch between synthetic and real audio data. To overcome the distribution mismatch, several semi-supervised learning approaches were proposed.

However, semi-supervised SED model generalization is inadequate for fitting the gap of distribution mismatch between synthetic and real audio data. In domain adaptation, the space of synthetic and real audio datasets can be treated as source and target domain respectively. The objective is transferring SED models trained on source domain to target domain.

2. PROPOSED METHOD

In this report, we propose an end-to-end domain adaptation method for robust SED under real scenarios. We employ a convolutional recurrent neural network (CRNN) as the backbone network for sound event detection. We propose a DACRNN network for joint learning of SED and domain adaptation (DA). We consider the impact of the distribution within a single sound on the generalization performance of the model by mitigating the impact of complex background noise on event detection and the self-correlation consistency regularization of clip-level sound event classification, these make the intra-domain of a single sound smoother. For cross-domain adaptation, adversarial learning through feature extraction network with weighted frame-level domain discriminator. Moreover, mixup and SpecAugment are applied in our system.

2.1. Network architecture

Convolutional recurrent neural network, which consists of several cascaded convolutional layers and gated recurrent units, is the representative model for sound event detection. The overall framework is shown in Figure 1. We employ a CRNN with 13 convolutional layers and 2 bidirectional gated recurrent units (Bi-GRU) as backbone feature extraction network. We use two types CNN blocks. In the first block (low-rise CNN Block), we use a layer normalization (LN) operation to replace commonly used batch normalization (BN) operation in the early stage of CRNN. In the second block (high-rise CNN Block), a shortcut is added between the first and last ReLU.

2.2. Self-correlation consistency regularization

In the previous sound event detection methods based on CRNN, researchers hope to improve the performance of sound event detection by combining the advantages of CNN and RNN in describing the local features and sequence features of samples. However, the CRNN sound event detection model trained by minimizing the classification cross entropy loss function in an end-to-end manner does not improve the insufficient ability of CNN structure to extract audio context information.

In order to overcome the shortcoming of CRNN sound event detection method, we propose an audio tag consistency constrained sound event detection method. The purpose method is to use CRNN network to improve the representation ability of CNN structure to audio sample context information by adding audio tag consistency constraints.

2.3. Weighted frame-level domain discriminator

Sound event detection is a frame level classification task. Therefore, the domain adaptation task of sound event detection should also act on frame level features. However, these frame level features are difficult to ensure the overall transfer adaptation of CRNN network to audio. How to introduce time structure information into the domain adaptation of frame level features is a problem we should consider.

$$DA_{loss-w} = \frac{1}{T}\sum_{t=1}^{T}(W_t BCE(f_t, d_t)) \quad (1)$$

$$W_t = |y_t^{stu} - y_t^{tea}|_2 \quad (2)$$

y_t^{stu} and y_t^{tea} represent the frame level output of student model and teacher model respectively. In other words, we use the average teacher model to adaptively weight the adversarial domain discriminator.

2.4. Data augmentation

Three data augmentations are applied in our system, namely random noise, spec-augmentation [2] and mixup [3].

In random noise, Gaussian noise is added to the spectrum as a small disturbance. Moreover, spectrum and noise spectrum are randomly send to teacher model and student model.

In mixup, all data is used to generate interpolated data. For weak-labeled and unlabeled clip, we use post processed teacher model prediction to generate interpolated label.

2.5. Post process

In model training, binarization, class-wise median filtering and tagging embedding are used in our system.

Binarization is used to teacher model output to get discrete prediction.

In Class-wise median filtering, we calculate the duration of each class in synthetic dataset to determine the window length of the median filter of each class.

In tagging embedding, we use audio tagging to adjust sound even detection. Binarization is also used on audio tagging.

$$p'_{tag} = \begin{cases} 0\ , & if\ p_{tag} \geq 0.5 \\ 1\ , & else \end{cases} \quad (3)$$

$$p'_{sed} = p_{sed} \odot p'_{tag} \quad (4)$$

2.6. Temperature process

The concept of temperature parameter T in sound event detection was first proposed by [4], which is used to soften the softmax output in Knowledge distillation field. In the field of sound event detection, temperature parameter T in the sigmoid function to soften the detection output only during model inference stage.

$$y_i = Sigmoid(z_i/T) = \frac{1}{1+\exp(-z_i/T)} \quad (5)$$

2.7. Utilizing weak prediction

By converting the sound event detection model into an audio tag model and using weak predictions and sets timestamp equal to the entire duration of the audio clip [5] has been proved to be beneficial to scenario 2 for PSDS2 score.

Different weak supervision pooling methods have slight differences in the performance of SED model. We found that using att pooling in the training stage and liner softmax pooling in the test inference stage can better transform SED model into audio tagging model.

3. EXPERIMENTS

All experiments are conducted on the DCASE 2022 domestic environment sound event detection (DESED) dataset, including *alarm/bell/ringing, blender, cat, dishes, dog, electric shaver/toothbrush, frying, running water, speech and vacuum cleaner*. The dataset includes 10000 synthetic audio clips and 19153 real audio chips in total. The synthetic audio clips are generated with Scaper. The real audio clips are extracted from Audioset, which contains 1576 weak labeled audio clips, 14388 unlabeled audio clips and 3189 strongly labeled audio clips.

In our experiments, we use the original audio sampling rate 44100HZ, and mel-spectrogram features are used as the basic features for sound event detection. Each audio clip in the dataset is transformed using fast Fourier transform with a 5646 points Hanning window and 707 hop length. Then, a mel filter-bank with 128 bandpass filters is applied to obtain the mel-spectrogram feature of the clip. As a result, a 10 second audio clip is converted into a (624,128) two-dimensional spectrogram. During training process, we use Adam to optimize all the loss functions of our approach, with a maximum learning rate of 0.002 for SED model and 0.001 for discriminator, and a learning rate rampup during the first 20 epochs.

3.1. Experimental results

In this section, we conduct experiments on validation set and public evaluation set to verify the effectiveness of our method.

Table 1 show the class wise metrics on all categories. The overall F1 scores of our model achieve 0.585 and 0.629 for validation dataset and public-evaluation dataset. The overall ER scores of our model achieve 0.76 and 0.66 for validation dataset and public-evaluation dataset.

Table 2 show the PSDS score and the influence about model ensemble. Our model achieve 0.421 and 0.521 PSDS1 score for validation dataset and public-evaluation dataset in the case of single model. Also as 0.840 and 0.738 PSDS2 score. And we can clearly see that model integration can effectively improve the detection performance.

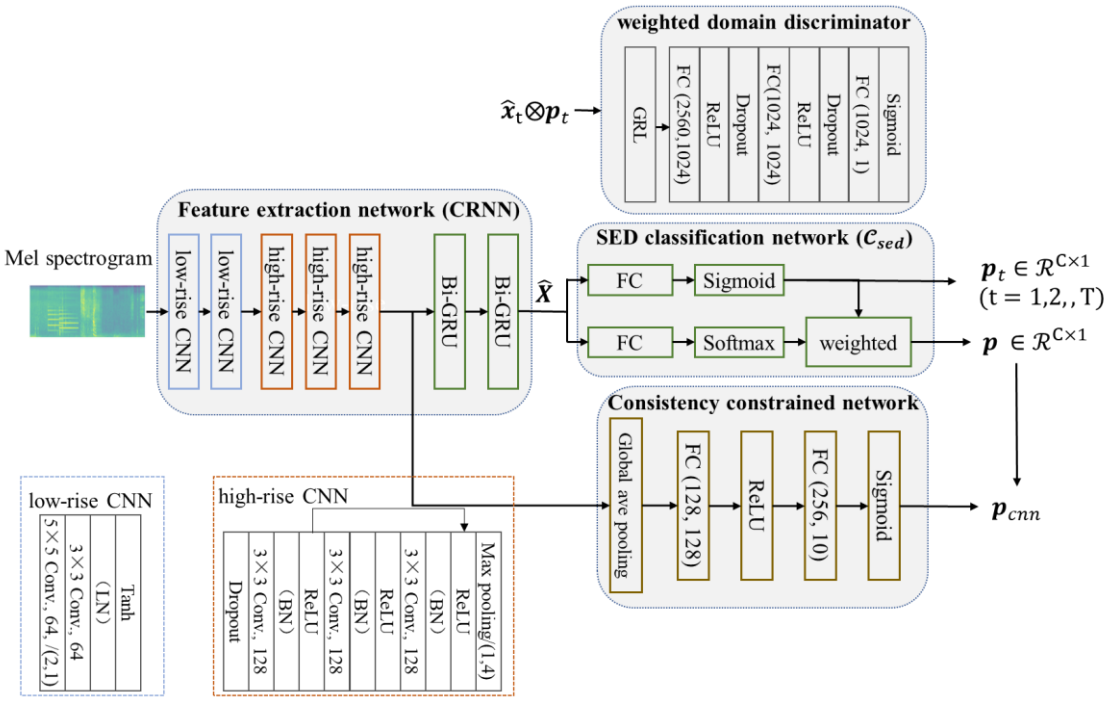

Figure 1: Framework of our SED system.

Table 1: SED performances of our model.

	validation set						public-evaluation set					
	F1	P	R	ER	Del	Ins	F1	P	R	ER	Del	Ins
alarm	0.513	0.596	0.450	0.85	0.55	0.30	0.631	0.712	0.566	0.66	0.43	0.23
blender	0.636	0.683	0.596	0.68	0.40	0.28	0.658	0.718	0.607	0.63	0.39	0.24
cat	0.555	0.596	0.519	0.83	0.48	0.35	0.827	0.847	0.808	0.34	0.19	0.15
dishes	0.387	0.468	0.329	1.04	0.67	0.37	0.463	0.584	0.383	0.89	0.62	0.27
dog	0.381	0.424	0.346	1.12	0.65	0.47	0.531	0.590	0.483	0.85	0.52	0.34
electric	0.800	0.833	0.769	0.38	0.23	0.15	0.598	0.724	0.509	0.69	0.49	0.19
frying	0.612	0.629	0.596	0.76	0.40	0.35	0.728	0.713	0.744	0.56	0.26	0.30
running water	0.520	0.571	0.477	0.88	0.52	0.36	0.441	0.574	0.358	0.91	0.64	0.27
speech	0.604	0.629	0.580	0.76	0.42	0.34	0.665	0.708	0.627	0.63	0.37	0.26
vacuum	0.841	0.881	0.804	0.30	0.20	0.11	0.753	0.817	0.698	0.46	0.30	0.16
average	0.585	0.631	0.547	0.76	0.45	0.31	0.630	0.699	0.578	0.662	0.421	0.241

Table 2: PSDS score for our submitted system.

	PSDS1		PSDS2	
model count	validation	public-evaluation	validation	public-evaluation
1	0.421	0.521	0.840	0.738
3	0.443	0.546	0.853	0.836
4	0.448	0.553	\	\

4. ACKNOWLEDGMENT

Thank UNISOC for providing computing device support.

5. REFERENCES

[1] http://dcase.community/workshop2022/.

[2] 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

[3] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "MixUp: Beyond empirical risk minimization," 6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc., pp. 1–13, 2018.

[4] X. Zheng.H. Chen, Y. Song. "Zheng USTC team's submission for dcase2021 task4 - semi-supervisedsound event detection" DCASE 2021.

[5] H. Nam, B.Y. Ko, G.T. Lee. "Heavily Augmented Sound Event Detection utilizing Weak Predictions" DCASE 2021.

[6] N. Turpault, R. Serizel, A. P. 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.

[7] F. Ronchini, R. Serizel, N. Turpault, and S. Cornell. The impact of non-target events in synthetic soundscapes for sound event detection. In Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021), 115–119. Barcelona, Spain, November 2021.