ADAPTIVE FOCAL LOSS WITH DATA AUGMENTATION FOR SEMI-SUPERVISED SOUND EVENT DETECTION

Technical Report

Yunhao Liang, Tiantian Tang, Yanhua Long

Shanghai Normal University, Shanghai, China
winnerahao@163.com, 1000479042@smail.shnu.edu.cn, yanhua@shnu.edu.cn

ABSTRACT

In this technical report, we describe our submission system for D-CASE2021 Task4: sound event detection and separation in domestic environments [1]. In our submissions, two different deep models are investigated. The first one is a mean-teacher model with convolutional recurrent neural network (CRNN). The second one is a joint framework with adaptive focal loss based on the Guided Learning architecture. To improve the performance of system, we propose to use various methods such as the specaugment data augmentation method, adaptive focal loss, event specific post-processing. To combine sound separation with sound event detection, we train models using the outputs of the sound separation baseline system. We demonstrate that the proposed method achieves the event-based macro F1 score of 44.4%, 0.428 in PSDS1 and 0.736 in PSDS2 on the validation set.

Index Terms— Acoustic event detection, Semi-supervised learning, Adaptive focal loss, Sound separation

1. INTRODUCTION

Sound event detection (SED) aims to detect and identify each sound event category and its onset and offset in the audio sequence. Recently, SED research includes audio event classification, abnormal sound detection, and automatic monitoring [2–4], etc. However, there are not many practical applications for sound event detection. Due to the diversity and complexity of real-life sound field environments, sound event detection can only be barely used in a few simple scenarios. The SED task requires a large amount of labeled training data, and these data cost a lot of cost for a large number of people to perform sound event categories and its onset and offset. In order to solve the problem of high cost of acquiring data labels in SED tasks, one solution is to use synthetic audio data to train the model. Current computer technology can synthesize high-quality audio sequences, and can generate a labeled synthetic audio dataset (such as DCASE2021 task4) for SED model training.

In this study, we introduce two systems for SED task. The first one is mean-teacher model (MT) [5], which is based on the official baseline system. The second one is a joint framework with adaptive focal loss based on the Guided Learning (GL) architecture [6, 7]. Moreover, two new methods are proposed to improve this system. To address the class imbalance of large-scale weakly labeled and unlabeled training data and different level of classification and detection difficulty of each target event, we propose a new training strategy with an adaptive focal loss together to enable an effective and more accurate model training. Furthermore, an event-specific post processing is designed to fix the prediction errors that result from outliers.

To incorporate sound separation (SS) into sound event detection, we fine-tune our SED models using the outputs from the official sound separation baseline system. Then, we fuse the event detection results of models trained by real data and separated data to get the final SS-SED ensemble system result.

2. METHODS

2.1. Network Architecture

2.1.1. Mean Teacher Model

The baseline system is inspired by the winning system from D-CASE 2018 Task 4 by Lu [5]. It uses a mean-teacher model which is a combination of two models: a student model and a teacher model (both have the same architecture). The student model is the final model used at inference time, while the teacher model is aimed at helping the student model during training and its weights are an exponential moving average of the student model's weights. To carry out the concept, the mean squared error between the outputs of the student model and the teacher model is added into loss function. And the network architecture is formed as a convolutional recurrent neural network (CRNN) [8], which consists of 7 layers of CNN blocks, 2 layers of bidirectional gated recurrent unit (GRU) cells, and an attention part for producing outputs. More details can be found in [9].

2.1.2. Guided Learning Model

The joint model architecture we proposed for both weakly supervised AT and AED consists of three parts: the teacher model, student model and event-specific post processing module. Both teacher and student models are also Convolution Recurrent Neural Networks (CRNN), but with different number of CNN blocks. The teacher model has five double-layer CNN blocks with a larger time compression scale that professional for a better audio tagging, while the student model only has three single-layer CNN blocks with no temporal compression scale for a better event boundary detection. This joint model is inspired by the Guided Learning (GL) model in [7]. But it is different from both the GL and traditional CRNN frameworks, besides the proposed deep feature distillation, the event-specific post processing and the two-stage model training strategy with adaptive focal loss, we also divide the AED and AT tasks into two independent branches. More details can be found in [6].

2.2. Data Augmentation

2.2.1. Mixup

Mixup [10] can improve the performance of deep neural network in many machine learning tasks by smoothing the distribution of samples in the feature space. This method creates a new data by interpolate between two raw data, while the labels are interpolated in the same way. The mixup smoothes out the decision boundary by adding pseudo data generated by mixing different data points $(x1; x2)$ and the corresponding labels $(y1; y2)$. The mixup is formulated as follows:

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda)\, x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda)\, y_j \end{aligned} \tag{1}$$

where x_i and x_j is two random chosen features, y_i and y_i is corresponding label respectively. λ is a random variable which follows the beta distribution. In our system, we used α =0.2

2.2.2. SpecAugment

SpecAugment [11] is an effective approach which shows significant performance improvement in acoustic speech recognition recently. It replaces values by zeros in randomly chosen time-frequency bands. SpecAugment includes three data augmentation methods, time warping, masking blocks of frequency channels, and masking blocks of time steps. In our method, we use frequency masks and time warping for data augmentation. All augmentations of SpecAugment are directly operated on the log-Mel filterbanks, which can save a lot of calculation time.

2.3. Combination of sound separation and SED

We also investigate to use the sound separation outputs to fine-tune the SED model for improving system performances. Instead of jointly End-to-End training of SED and separation system, we choose to directly use the pre-trained official sound separation baseline [12, 13] to perform the sound separation for all the training, development and evaluation data. Specifically, we use all of the separated clips to fine-tune our SED model (with the batch normalization layer weights were kept frozen). The proposed approach is significantly different from the one that used in the challenge official sound separation baseline. In the official baseline separation system, it has 8-channel separated sources, but for the real SED dataset, a single audio clip even contains less than 8 events. So the separation results of the baseline system will include background events. Using all the separated sources to train the SED model can potentially introduce a mismatch. So we use a pre-trained SED classifier to pre-classify the 8-channel separated sources of each input audio to obtain their prediction labels. These labels are then used to fine-tune the MT SED system.

2.4. Event-specific post processing (ESP)

Median filtering (MF) has proved effective in smoothing the noisy outputs of the student model for AED tasks [9]. Conventional MF with fixed window size is no longer suitable for this task. Recent works in [7, 14] used group of median filters with adaptive window size by calculating the average duration of events with strong labels on the development set. However, the duration of each target event is not an uniform distribution, using the average event duration to optimize the MF window size may not the optimal. So we propose to use event-specific MF window size as:

$$\mathcal{W}_c = \left(\frac{1}{N_c} \sum_{i=1}^{N_c} L_i \right) \cdot \eta \tag{2}$$

where $\mathcal{W}_c, c = 1, 2, ..., C$ is the MF window size of event class c, N_c is the segment index for the inflection point of cumulative distribution of short-to-long sorted segments of c-class target event. L_i is duration of i-th segment for event c. η is a scaling factor and set to $1/3$ in our experiments. All the training clips with strong labels are used to compute $\mathcal{W}_c$.

2.5. Adaptive focal loss

Motivated by the principle of focal loss in [15], here we aim to improve the model training by combining the above BCE loss with an adaptive focal loss that defined as follows:

$$\mathcal{L}_{af} = -\frac{1}{CK} \sum_{j=1}^{C} \sum_{i=1}^{K} \left(1 - p_{ij}^{\gamma}\right) \cdot \log\left(p_{ij}\right) \tag{3}$$

where γ is a scaling factor to control the loss contribution of posterior probability p_{ij} for i-th clip, j-th target-event category. K is the total size of audio clips with both weakly and strong labels in a minbatch, C is the number of target-event categories.

3. EXPERIMENTS

3.1. Datasets and Features

The training set of our SED system contains a weakly-labeled training set (1578 clips), an unlabeled training set (14412 clips), and a synthetic strongly labeled set (10000 clips). The validation set contains 1168 strongly-labeled clips. The public test set contains 692 strong-labeled clips. We extract 128-dimensional log-Mel filterbanks from the input audio. The window size and the hop size are 2,048 points and 256 points, respectively, in 16 kHz sampling. We fix the length of the feature sequence to 625 frames (corresponding to around 10 seconds). To make the length of feature sequences the same, we perform zero-padding for shorter sequences and truncation for longer sequences from their last frames. Then, we perform the normalization to make the feature sequences have zero means and unit variances over the training data.

3.2. Setup

For the data augmentation, we apply 25% mix operation in a mini-batch for mixup method and apply 50% operation in a mini-batch for sepcaugment method in our system. For the SS-SED system, we use the baseline system of sound separation to separate the training set of the SED.

4. RESULTS

All techniques are examined on DCASE 2021 Task4 validation set and results are shown in Table 1. 'SED-Baseline' and 'SS+SED-baseline' are two official baselines. 'MT' and 'GL' are our baseline models without the proposed ESP, specaugment(Spec), adaptive focal loss(AFL) and the sound separation(SS).

Table 1: $F1$-scores (%) and PSDS metrics of the proposed methods.

ID	Method	PSDS1	PSDS2	Collar-based F1(%)
0	SED-Baseline	0.342	0.527	40.1
1	SS+SED-baseline	0.373	0.549	44.3
2	MT+ESP	0.353	0.569	40.6
3	GL+ESP	0.263	0.531	38.1
4	MT+ESP+Spec	0.397	0.640	41.4
5	GL+ESP+Spec	0.281	0.555	38.5
6	MT+ESP+Spec+AFL	0.418	0.717	42.0
7	GL+ESP+Spec+AFL	0.328	0.575	41.0
8	MT+ESP+Spec+AFL+SS	0.428	0.736	44.4

In Table 1, we see that the last method with all the proposed techniques achieves the best results for both AED and AT tasks, it outperforms the 'SS+SED-baseline' system significantly by absolute 5.5% and 18.7% in PSDS1 and PSDS2. However, the 'GL+ESP+Spec+AFL' system which we proposed based on GL only achieved 4.8% improvement in PSDS2.

5. CONCLUSION

In this technical report, we have described the techniques that used in our submission systems for DCASE2021 Task4. Our system has been developed based on two model architecture, including the Mean Teacher model and the Guided Learning model. The data augmentation techniques, the event-specific post-processing, adaptive focal loss, and the sound separation are also used to further improve system performances. Experimental results on the validation set demonstrate that these techniques are helpful for improving the sound event detection performance, and the Mean teacher model significantly outperforms the baseline. Unfortunately, the proposed Guided Learning system only achieved small improvement over the baseline. Investigating the class-wise performance more carefully and generalizing the proposals to other acoustic event detection tasks to develop more effective technique is our future work.

6. REFERENCES

[1] http://dcase.community/challenge2021/.

[2] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 81–94, 2016.

[3] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[4] Q. Jin, P. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based video retrieval using audio," in *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2012, pp. 2085–2088.

[5] L. JiaKai, "Mean teacher convolution system for DCASE 2018 task 4," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, 2018.

[6] Y. Liang, Y. Long, Y. Li, and J. Liang, "Joint weakly supervised at and aed using deep feature distillation and adaptive focal loss," 2021.

[7] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning for weakly-labeled semi-supervised sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 626–630.

[8] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[9] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019.

[10] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2017.

[11] 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." in *Interspeech 2019*, 2019, pp. 2613–2617.

[12] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, "Universal sound separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2019, pp. 175–179.

[13] 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.

[14] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 task 4," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.