AN EFFECTIVE CONSISTENCY REGULARIZATION TRAINING BASED MEAN TEACHER METHOD FOR SOUND EVENT DETECTION

Technical Report

Yunlong Li[1,2]*, Ying Hu*[1,2]*, Xiujuan Zhu*[1,2]*, Yin Xie*[1,2]*, Shijing Hou*[1,2]*, Liusong Wang*[1,2]*, Zihao Chen*[1,2]*, Mingyu Wang*[1,2]*, Wenjie Fang*[1,2]*,*

[1] Xinjiang University, School of Information Science and Engineering, Urumqi, China
[2] Key Laboratory of Signal Detection and Processing in Xinjiang, Urumqi, China
{liyunlong, huying}@stu.xju.edu.cn

ABSTRACT

This technical report describes the system we submitted to DCASE2021 Task4: Sound Event Detection in Domestic Environments. Specifically, we apply three main techniques to improve the performance of the official baseline system. Firstly, to improve the detection and classification ability of the CRNN model, we propose to add an auxiliary branch to the CRNN network. Consistency loss of mean teacher method is improved by auxiliary branch. Secondly, we propose to add an MDTC module to the CRNN network so that the receptive fields of the network can be adjusted according to the short-term and long-term correlation. Thirdly, several data-augmentation strategies are adopted to improve the generalization capability of the network. Experiments on the DCASE2022 Task4 validation dataset demonstrate the effectiveness of the techniques used in our system. As a result, the best PSDS1 is 0.408 and the best PSDS2 is0.754.

Index Terms— Sound Event detection, Semi-supervised learning, mean teacher, consistency loss

1. INTRODUCTION

Sound Event Detection(SED) is a task that detects both the onset and offset of sound events and identifies event categories. Our system uses CRNN [2] network with mean teacher [3] semi-supervised learning method based on the official baseline system [4] [5]. We implement the following methods to improve the network performance:

- By adding auxiliary branches into the CRNN network, the consistency criterion of mean-teacher model training is extended by adding additional consistency loss.
- By adding a Multi-dilation time convolution(MDTC) module to the main branch of CRNN, sound events' short-term and long-term correlations can be modeled by aggregating features of different time scales.
- Mixup [6], FilterAugment [7] and Cutout [8] data-augmentation strategies are used to improve the generalization capability of the detection system.

2. PROPOSED METHODS

2.1. model

Our network structure is based on the CRNN network of the Baseline system. The feature extractor of CRNN is a stack of 7 convolutional layers. The kernel size of each convolutional layer is (3,3). Each convolution block is followed by a gaussian error linear unit (GeLU) [10] activation and batch normalization (BN) [11]. Average pooling is performed after each block, 4-times reduce the output time resolution of the CRNN model, and the frequency axis is pooled to 1. Then the proposed MDTC module is fed into the feature extractor, and its output is fed into the bi-directional gated recurrent unit(Bi-GRU), fully connected layer and Sigmoid to get a strong prediction and then a weak prediction of 10 acoustic events are obtained by Linear Softmax.

2.2. Auxiliary branche

Inspired by clip-level consistency training [12], an auxiliary branch is introduced after the feature extractor to improve the feature representation ability and classification generalization ability of the CRNN network. This branch comprises BI-GRU and classifier and only computes the consistency loss with the CRNN main branch. Therefore, the total loss consists of strong prediction loss, weak prediction loss, the consistency loss of mean teachers, and the consistency loss between the main branch with the auxiliary branch output.

2.3. MDTC module

After feature extraction by CNN, the frequency dimension is sampled to 1, so a feature is obtained. Inspired by the time convolution network (TCN) in ConvT-Tasnet [13], we propose an MDTC module consisting of cascaded CNN blocks. The expansion factor d of CNN blocks increases exponentially, which increases the temporal receptive field. Finally, the output features of each CNN block are aggregated together through the aggregator to obtain the total output.

2.4. Data augmentation

For all training data, including weakly labeled data, unlabeled data and synthesized strong labeled data, we use Mixup [6], FilterAugment [7] and Cutout [8] methods to generate augmented data. The

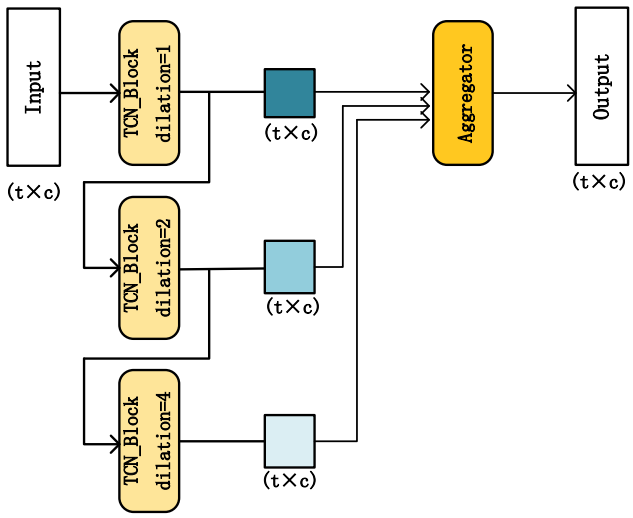

Figure 1: MDTC module

Mixup method generates augmented data by getting the weighted sum of the two pieces of data. For FilterAugment, it applies different weights in different frequency bands to generate augmented data. For Cutout, it generates augmented data by randomly masking random square areas of time-frequency features.

3. EXPERIMENTAL EVALUATION

3.1. Dataset

We conduct experimental evaluations on the DCASE2022 Task4 dataset. The dataset contains 1578 audio clips with weak labels, 10000 synthesized audio clips with strong labels and 14412 unlabeled audio clips. In addition, we used 3470 real audio clips as external data to train.

3.2. Experimental setup

We choose the Adam optimizer with a learning rate of 0.001, and the total training epoch is 200. Each 10-second audio clip is resampled to 16 kHz. The log-mel spectrogram uses 2048 STFT windows with a hop size of 255 and 128 Mel-scale filters, so the size of the input features is 628×128. All experiments were conducted on a GeForce RTX TITAN GPU 24GB RAM.

3.3. Experimental results

The model we submitted is shown in Table 1. We report the results of energy consumption (kWh) obtained by Codecaron and the two main challenge metrics, PSDS-1 and PSDS-2. None of the four models adopt the model fusion strategy. Model1 and model2 use attentional layers and average pooling to generate weak predictions for auxiliary branches. And all of the above techniques are used simultaneously. Model3 and Model4 have the same network structure as model1 and model2, respectively, but they use an additional 3470 external data to train.

The experimental results show that using the attentional layer to generate weak prediction in the auxiliary branch has the best performance for PSDS1, and using average pooling has the best performance for PSDS2. Our overall PSDS(PSDS1 + PSDS2) was 1.162, 29% higher than baseline.

Method	train(kwh)	test(kwh)	psds1	psds2
baseline	1.717	0.030	0.336	0.536
model1	2.718	0.017	**0.408**	0.607
model2	3.771	0.006	0.095	**0.754**
Baseline(external)	2.418	0.027	0.351	0.552
model3(external)	3.791	0.010	**0.398**	0.640
model4(external)	3.317	0.015	0.215	**0.735**

Table 1: Final results of the models submitted

4. REFERENCES

[1] https://dcase.community/challenge2022/task-sound-event-detection-in-domestic-environments.

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

[3] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for dcase 2019 task 4," *Orange Labs Lannion, France, Tech. Rep*, 2019.

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

[5] 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)*, Barcelona, Spain, November 2021, pp. 115–119.

[6] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[7] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugment: An acoustic environmental data augmentation method," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4308–4312.

[8] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[9] 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.

[10] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[11] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[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] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.