A EFFICIENCE SOUND EVENT DETECTION SYSTEM FOR DCASE 2024 TASK 4

Technical Report

ZunXue Niu[1,2], *Ying Hu*[1,2], *Xin Fan*[1,2], *Jie Liu*[1,2], *Ye Dong*[1,2], *Fujie Xu*[1,2], *ShangKun Tu*[1,2], *KaiMin Cao*[1,2], *JiaBo Jing*[1,2], *Qiong Wu*[1,2], *QingJing Wan*[1,2],

[1] XinJiang University, School of Information Science and Engineering, Urumqi, China
{niuz2267}@stu.xju.edu.cn
[2] Key Laboratory of Signal Detection and Processing in Xinjiang, Urumqi, China

ABSTRACT

This technical report describes the system we submitted to DCASE2024 Task4: Sound Event Detection with Heterogeneous Training Dataset and Potentially Missing Labels. Specifically, we apply three main techniques to improve the performance of the official baseline system. Firstly, We exploiting a dual-branch convolutional recurrent neural network (CRNN) structure including the main branch and auxiliary branch. We adopt an SCT strategy to apply the self-consistency regularization in addition to the Mean Teacher loss to maintain the consistency between the outputs of the auxiliary and main branches. Secondly, a HTA module is designed to aggregate the information at different temporal resolutions 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 robust of the network. Experiments on the DCASE2024 Task4 validation dataset demonstrate the effectiveness of the techniques used in our system.

Index Terms— Sound Event Detection, Heterogeneous Dataset, data augmentation, consistency regularization

1. INTRODUCTION

Sound event detection (SED) is the task of detecting the categories of sound events and the timestamps of their corresponding occurrence [1]. In this report, we utilize a self-consistency training (SCT) strategy for semi-supervised SED, this method adopts a dual-branch CRNN [2] structure, including the main branch and auxiliary branch. The auxiliary branch assists the main branch in the form of consistency regularization to train a model with better generalization performance [3]. Specifically, We implement the following methods to improve the network performance:

(i) We propose a self-consistency training (SCT) strategy that by adding auxiliary branches into the CRNN network and apply self-consistency regularization in addition to the Mean Teacher [4] loss.

(ii) A hierarchical temporal aggregation (HTA) module is designed to aggregate the features of different temporal resolutions, which are added to the main branch of CRNN, so sound events' short-term and long-term correlations can be modeled by aggregating features of different time scales.

(iii) We utilized the Mixup [5], SpecAugment [6], Audio cutmix [7] [8], RandomLinearFader (RLF) [9] data augmentation to improve the generalization capability of the detection system.

2. METHODS

Our network structure is based on the CRNN network of the Baseline system [10] [11]. The feature extractor of CRNN is a stack of 7 convolution layers and we adopt the FDY [12] instead of the traditional convolution. The kernel size of each convolution layer is(3,3). Each convolution block is followed by a gaussian error linear unit (GeLU) [13] activation and batch normalization (BN) [14]. 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 HTA module is followed by the feature extractor in the main branch, 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, and the feature extractor output in the auxiliary branch is directly fed into Bi-GRU, and and the weak prediction is finally obtained by attention Softmax. And we also design a Fusion module [15] to combine the latent feature from extractor with embedding from pre-trained model Beats as Fig.1 shown.

Inspired by clip-level consistency training [16], 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.

After feature extraction by FDY-CNN, the frequency dimension is down-sampled to 1, so a feature is obtained. Inspired by the time convolution network (TCN) in ConvT-Tasnet [17], we propose an HTA module consisting of cascaded hierarchical TCN blocks. The dilation factor d of TCN blocks increases exponentially, which increases the temporal receptive field. Finally, the output features of each TCN block are aggregated together through the aggregator to obtain the total output.

3. DATA AUGMENTATION

We adopted the Mixup [5], SpecAugment [6], Audio cutmix [7] [8], RLF [9] data augmentation methods to generate augmented data. The Mixup method generates augmented data by getting the weighted sum of the two pieces of data. While SpecAugment ran-

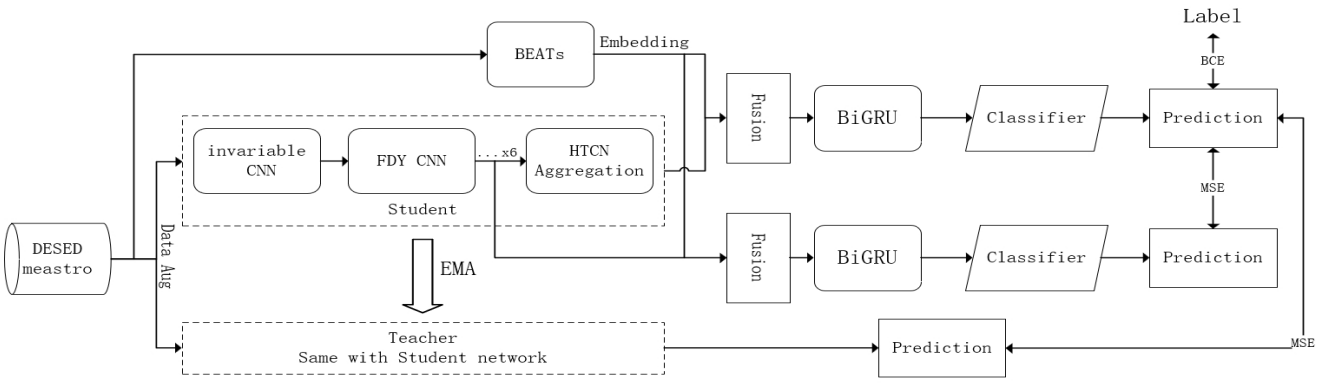

Figure 1: Sound Event Detection pipeline.

domly mask out a continuous segment of the spectrum along the time or frequency axis. Audio cutmix can alter the boundaries of sound events and contextual information, encouraging the model to learn more robust feature representations with limited context. RLF involves randomly applying linear volume changes throughout the entire audio clip to simulate variations in the position of sound sources.

4. EXPERIMENTAL

4.1. Dataset

We conduct experimental evaluations on the DCASE2024 Task4 dataset[1]. The dataset cosistt of DESED dataset and Meastro Real, where the DESED contains 1578 audio clips with weak labels, 10000 synthesized audio clips with strong labels and 3470 real audio clips with strong labels as well as 14412 unlabeled audio clips, and Meastro Real contains real-life recordings with a length of approximately 3 minutes each, recorded in a few different acoustic scenes, the audio was annotated using Amazon Mechanical Turk, with a procedure that allows estimating soft labels from multiple annotator opinions.

4.2. Experimental setup

We choose the Adam optimizer with a learning rate of 0.0026, and the total training epoch is 200. Each audio clip is resampled to 16 kHz. The log-mel spectrogram uses 2048 STFT windows with a hop size of 256 and 128 Mel-scale filters, so the size of the input features is 626×128. This features are normalized to zero mean and unit variance before being fed into the network. The batch size is set to 60. Each batch consists of 12 soft-labeled, 12 strongly-labeled include 6 synthetic and 6 real strongly labeled, 12 weakly labeled, as well as 24 unlabeled audio clips.All experiments were conducted on a GeForce RTX TITAN GPU 24GB RAM.

4.3. Experimental result

The system's result which we submitted is shown in Table 1, and the Energy Consumption is shown in Table 2. The experimental results show that our proposed method outperformance the baseline on PSDS1 and PSDS1 (sed score).

[1]https://dcase.community/challenge2024/task-sound-event-detection-with-heterogeneous-training-dataset-and-potentially-missing-labels

Table 1: The results of the system submitted.

Methods	PSDS1	PSDS1 (sed score)	mean pAUC
Baseline 2024	0.480	0.490	**0.730**
ours' system	**0.493**	**0.532**	0.657

Table 2: The Energy Consumption of system submitted.

Methods	train(kwh)	test(kwh)
Baseline 2024	1.666	0.143
ours' system	3.273	0.062

5. REFERENCES

[1] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine (SPM)*, vol. 38, no. 5, pp. 67–83, 2021.

[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] 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.

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

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

[6] 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.

[7] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with lo-

calizable features," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2019, pp. 6023–6032.

[8] Y. Li, X. Wang, H. Liu, R. Tao, L. Yan, and K. Ouchi, "Semi-supervised sound event detection with local and global consistency regularization," *arXiv preprint arXiv:2309.08355 (Accepted by ICASSP 2024)*, 2023.

[9] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Byol for audio: Exploring pre-trained general-purpose audio representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 31, pp. 137–151, 2022.

[10] 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*, 2019.

[11] F. Ronchini, R. Serizel, N. Turpault, and S. Cornell, "The impact of non-target events in synthetic soundscapes for sound event detection," *arXiv preprint arXiv:2109.14061*, 2021.

[12] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency Dynamic Convolution: Frequency-Adaptive Pattern Recognition for Sound Event Detection," in *Proc. Interspeech 2022*, 2022, pp. 2763–2767.

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

[14] 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.

[15] Y. Hu, Y. Chen, W. Yang, L. He, and H. Huang, "Hierarchic temporal convolutional network with cross-domain encoder for music source separation," *IEEE Signal Processing Letters*, vol. 29, pp. 1517–1521, 2022.

[16] L. Yang, J. Hao, Z. Hou, and W. Peng, "Two-stage domain adaptation for sound event detection." in *DCASE*, 2020, pp. 230–234.

[17] 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.