SOUND EVENT DETECTION SYSTEM USING PRE-TRAINED MODEL FOR DCASE 2023 TASK 4

Technical Report

Wei-Yu Chen, Chung-Li Lu, Hsiang-Feng Chuang, Yu-Han Cheng, Bo-Cheng Chan

Chunghwa Telecom Laboratories, Taiwan
{weiweichen, chungli, gotop, henacheng, cbc}@cht.com.tw

ABSTRACT

In this technical report, we briefly describe the system we designed for Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 Challenge Task4: Sound Event Detection with Weak Labels and Synthetic Soundscapes. Our best single system combines the embedding obtained by VGGSK and BEATs, using GRU to classify sound events for each frame. Thresholding and smoothing are utilized during the post-processing stage. The mean teacher method is applied for semi-supervised learning with the EMA strategy to update parameters of the teacher model. To utilize unlabeled data, pseudo label is generated by the student model. In the process of data augmentation, we utilize techniques such as mix-up, Gaussian noise and embedding masking. The submitted single system trained with extra data achieves the PSDS1 of 0.529 and the PSDS2 of 0.78 on the validation set.

***Index Terms*—**DCASE, sound event detection, mean teacher, pre-training, consistency training

1. INTRODUCTION

The goal of sound event detection is to know which sound events occurred in the sound clip as well as the start time and end time of the sound event after a sound comes in. One difficulty of using supervised learning for sound event detection is that the cost of obtaining and marking these sound events after collecting sound data is considerable, and it can cause bias easily from different annotators. Therefore, the audio files of the DESED data set are obtained by two methods: real ambient sound and synthetic sound. The labeled data is divided into three types, namely strong label, weak label and unlabeled. In addition to using DESED to provide data sets, participants are encouraged to use external dataset or pre-trained embedding. However, it is still necessary to submit a result that does not use external dataset and pre-trained embedding. In addition, at least one of the submitted systems must not use ensemble.

For the part that can use external dataset or pre-trained embedding, Bidirectional Encoder representation from Audio Transformers (BEATs) [1] based on Transformer are used to obtain embedding features. The model only trains VGGSK [2] based on CNN. Therefore, Exponential Moving Average (EMA) strategy only updates the VGGSK and Gated Recurrent Unit (GRU) parts in the process of updating the teacher model from the student model. For VGGSK input data, mix-up, Gaussian noise and ICT are used for data augmentation, and mask embedding is done on BEATs to obtain embedding features. The main purpose is to improve the prediction results without increasing the inference time.

In this report, we describe our submission system for DCASE 2023 Task4. The content includes network architecture, data augmentation method and the fusion strategy and post-processing method of ensemble after obtaining each trained model.

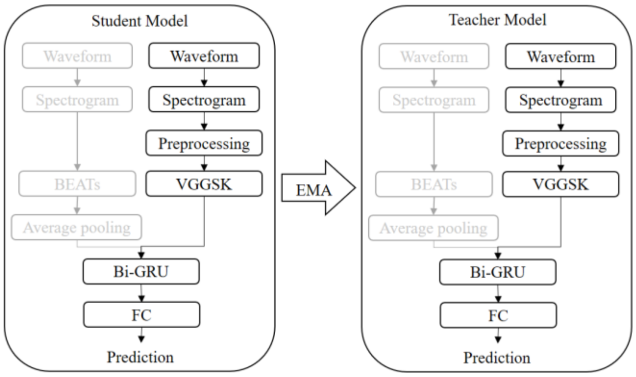

Figure 1: The proposed sound event detection system structure.

2. PROPOSED METHODS

2.1. Network architecture

The network architecture mainly adopted VGGSK and BEATs, in which data augmentation is performed in the pre-processing part of VGGSK. We obtain the embedding from the BEATs model and apply masking to the embedding representation. The entire architecture can be seen from the student model in Figure 1. The VGGSK model in the network architecture is composed of VGG block [3] and four residual blocks with selective kernels (SK) [4]. In the BEATs part, the official pre-trained model of BEATs$_{\text{iter3+}}$ is used. For the supervised learning part, Binary Cross-Entropy (BCE) is used as loss function to calculate the loss.

2.2. Semi-supervised learning

As to train unlabeled data, the mean-teacher in semi-supervised learning is used, in Figure 1. After the student model has been learned through supervised learning, the teacher model is updated using the EMA strategy. At this time, the unlabeled data is predicted by the student model, and then being used as the pseudo labels for the unlabeled data in teacher model. However, the loss between the student model and the teacher model is calculated by Mean Square Error (MSE). In addition, ICT is added to the loss function to improve the recognition results.

2.2.1. Interpolation Consistency Training

The interpolation consistency training (ICT) [5] is to perform interpolation calculation on the prediction results of the model, for the purpose of processing ambiguous samples and improving the generalization ability. We define the ICT loss function by

$$L_{ICT} = MSE\begin{pmatrix} \theta\left(\lambda d_i + (1-\lambda)\, d_j\right), \\ \lambda\theta'(d_i) + (1-\lambda)\,\theta'(d_j) \end{pmatrix} \tag{1}$$

where θ and θ' denote a student model and a teacher model, d_i and d_j denote data points, and λ is randomly sampled from a Beta distribution.

3. EXPERIMENTS

3.1. Single Model of Submitted Systems

Table 1 shows the performance of each stage of our single model in the submission system. The baseline model initially utilized both the CRNN and BEATs models [6]. However, it was modified by replacing the CRNN model with the VGGSK model. This change resulted in an improvement in PSDS1, increasing it from 0.500 to 0.518. After incorporating the strong real dataset into the training process, an improvement was observed in PSDS2, with the value increasing from 0.764 to 0.776. By incorporating the ICT method, significant improvements were achieved in the results. PSDS1 improved to 0.529, indicating a substantial enhancement, while PSDS2 saw a remarkable improvement to 0.780.

	PSDS1	PSDS2
CRNN+BEATs(Baseline)	0.500	0.762
VGGSK+BEATs	0.518	0.764
+Strong Real Dataset	0.516	0.776
+ICT	0.529	0.780

Table 2: The single model on validation set.

3.2. Results of Submitted Systems

Table 2 shows the performance of our submitted systems, all of which are based on the VGGSK model. System 2, 3 and 4 incorporate extra data by utilizing the BEATs model. System 1 utilizes the VGGSK model. The system 2 is described in section 3.1. The pool of methods to form system 3 and 4 were trained from identical model structure but with different data augmentation skills. We randomly get several candidate ensemble systems by simply averaging results from 6 methods in the pool of methods, and system 3 is chosen by highest PSDS1 0.552 among all candidate systems. Likewise, system 4, which is chosen by highest PSDS2 0.799, is an ensemble of 15 methods.

System	Extra data	PSDS1	PSDS2
CRNN (Baseline)		0.359	0.562
CRNN+BEATs(Baseline)	✓	0.500	0.762
System 1		0.424	0.633
System 2	✓	0.529	0.780
System 3	✓	0.552	0.794
System 4	✓	0.542	0.799

Table 2: The PSDS on validation set.

4. REFERENCES

[1] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, F. Wei , “BEATs: Audio Pre-Training with Acoustic Tokenizers,” arXiv preprint arXiv:2212.09058, 2022

[2] S. J. Huang, C. C. Liu, C. P. Chen, C. L. Lu, B. C. Chan, Y. H. Cheng, H. F. Chuang, “CHT+ NSYSU SOUND EVENT DETECTION SYSTEM WITH DIFFERENT KINDS OF PRETRAINED MODELS FOR DCASE 2022 TASK 4. ” 2021

[3] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in International Conference on Learning Representations, 2015.

[4] X. Li, W. Wang, X. Hu, and J. Yang, “Selective kernel networks,” In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 510-519), 2019

[5] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, “Interpolation consistency training for semi-supervised learning,” arXiv preprint arXiv:1903.03825, 2019.

[6] http://dcase.community/workshop2023/.