SRCB-BIT TEAM'S SUBMISSION FOR DCASE2022 TASK4

Technical Report

Liang Xu[1,2], *Lizhong Wang*[2], *Sijun Bi*[1], *Hanyue Liu*[1],
Jing Wang[1], *Shenghui Zhao*[1], *Yuxing Zheng*[2],

[1] School of Information and Electronics, Beijing Institute of Technology, Beijing, China
[2] Samsung Research China-Beijing (SRC-B), Beijing, China
{xuliang, 3120200743, 1120183530, wangjing, shzhao}@bit.edu.cn
{lz.wang, yxing.zheng}@samsung.com

ABSTRACT

In this technical report, we present our submitted system for DCASE2022 Task4: Sound Event Detection in Domestic Environments. We propose three main ways to improve the performance of the network. First, we use the frequency dynamic convolution (FDY) which applies kernel that adapts to frequency components of input to improve physical inconsistency in 2D convolution on sound event detection (SED). Then, we propose a weight raised temporal contrastive loss based coherence learning to improve the continuity of event prediction and the switching efficiency of event boundaries. Third, we use pre-trained model PANNS in this task and propose two methods to fuse the features from PANNs and our model which improve the PSDS1 and PSDS2 score respectively. The system we submitted is based on the mean-teacher architecture, and the PSDS1 and PSDS2 score on the development dataset can reach 0.482 and 0.835 respectively.

***Index Terms*—** Sound event detection, Semi-supervised learning, mean-teacher, pre-trained model, frequency dynamic convolution

1. INTRODUCTION

Sound Event Detection (SED) aims to detect sound event categories and their corresponding time of onset and offset (timestamp) in a sound clip. In order to adapt to this task better, a large amount of data with strong labels are required. However, since hand-labeling these collected data is extremely costly, the data sets are difficult to obtain. At the same time, this kind of sound should be collected in the real environment to be applied in real life. As an alternative, strongly labeled dataset are synthesized from foreground and background datasets. However, it is still difficult to obtain only a large foreground dataset containing with strong labels. Therefore, limited strong labeled data is trained by combining an example amount of weakly labeled data whose labels only include the sound event types without timestamps of the events and unlabeled dataset whose label has no information at all.

The DCASE 2022 Task 4 is the follow-up to DCASE 2021 Task 4. The task evaluates systems for the detection of sound events using weakly labeled data (without timestamps). The target of the systems is to provide not only the event class but also the event time localization given that multiple events can be present in an audio recording. Compared to DCASE 2021 Task 4, this year's data sets are the same as last year, while there is no source separation pre-processing. Besides, this year's task adds extra data resources including pre-trained models and allowed datasets. We find that the results of some of top-ranked models in the DCASE Challenge 2021 Task 4, were based on a mean-teacher model [1] trained mainly by both weakly labeled and unlabeled data with consistency regularization.

In this report, we propose an SED model based on mean-teacher model and then train it using all the training data, including strongly labeled, weakly labeled and unlabeled data. Next, the trained model is used to test the performance in the evaluation set. In order to better complete the task, we consider three different methods: 1) the frequency dynamic convolution (FDY) [2], 2) a new weight raised temporal contrastive loss function that acts as label, and 3) pre-trained models to obtain embedding features using PANNs [3].

2. PROPOSED METHODS

2.1. Network architecture with FDY convolution

Nam et al. demonstrated that the frequency domain energy distribution is different for different events in the SED task. The convolution neural network(CNN) commonly used in deep learning has translation invariant properties. Translation invariance means that the system will produce exactly the same response regardless of how its input is translated. In the field of computer version, the translation invariance of convolution means that objects appearing in the image can be correctly detected after translated to any position. But for SED, this property may misjudge two events with similar energy but different distributions, thus reducing the overall performance of the SED task. This means that simply increasing depth or width of CNN architecture cannot improve the detection ability.

Dynamic convolution enhances representation capability of CNN architecture by applying input-adaptive kernel on convolution layer. By extracting attention weights for the weighted sum of basis kernels, dynamic convolution generates appropriate kernel for given input. This means dynamic convolution can overcome the translation invariance of convolution. Here, we use the FDY convolution proposed by Nam et al. in [2] to replace the 2D Conv. in the baseline and we set the same hyperparameter for each FDY convolution block as in [2].

For CNN part, the first block is constructed by a 2D CNN block, then followed by 6 FDY convolution blocks. Each FDY block contains a FDY convolution layer, a BatchNorm layer and an activate function. The number of filters for 6 FDY convolution blocks are [64, 128, 256, 256, 256, 256] respectively, and frequency pooling rate for each stage is set to 2, whereas the total temporal pooling rate is set to 4.

2.2. Network architecture with PANNs

Despite the attention pald to competitions such as DACSE, how well pre-trained audio pattern recognition systems perform on large-scale datasets is still an open question. PANNs is a pre-trained model for audio tagging trained on Audioset, and Kong proposed a Wavegram-Logmel-CNN system which achieves audio set tagging with mean precision (mAP) of 0.439 [3]. It is worth noting that although PANNs is trained on weakly labeled data, it can still show excellent performance when transferred to other strongly labeled tasks. Based on the above reasons, we fuse the CRNN network with the PANNs network and propose two fusion schemes to improve the scores of PSDS1 and PSDS2.

Here we use the embedding features (2048 dims) and the prediction framewise features (527 dims) output by PANNs to assist the SED task. First, considering that the prediction framewise features output by PANNs after transfering to SED task have temporal characteristics, we weighted add the output results of PANNs and the output of GRU respectively after feature dimensionality reduction.The pre-training features and the features extracted by CRNN are reduced to 10 dimensions, and the fused features are used as input to the classifier and the probability of each category on each frame is calculated. Second, we concatenate the embedding features of PANNs and the output of GRU in the feature dimension and send the concatenated features to the classifier for prediction. Before the fusion of these two methods, the frame length of the pre-trained features and GRU output features must be unified.

The first fusion method improves the frame-level accuracy according to the mutual constraints of the two features on the same frame, and the second fusion method improves the accuracy of category prediction by providing high-dimension features. In our submitted 4 systems, each system use at most one fusion method.

2.3. Weight raised temporal contrastive loss

In today's SED tasks, the loss functions of most models tend to use BCE and MSE. It is true that they are effective methods to improve the convergence of the model to make the label predicted by the model completely approach the real label, but the model cannot learn the independent and effective information of each event very well in this process.

BCE and MSE loss function effectively ignores the particular relevance of instances near event boundaries which are expected to facilitate boundary detection; while instances within events are expected to reflect more stationary or coherent behavior in feature representation [4]. Thus, Kothinti et al. proposed a loss function applied between the ground truth and the output of convolutional layer in CRNN to improve event boundary recognition[4].

However, the time information contained in the output features of the convolutional layer in CRNN is very weak, which make it impossible to effectively constrain the time series information, and the randomness of the model prediction result at the beginning of training is considered too strong. Therefore, We propose a temporal contrastive loss wr-TCL whose weights rise with training epochs and use this loss function to contrast the model together with BCE and MSE.

wr-TCL =

$$\sigma(-\alpha\sum_{i=1}^{N}\sum_{t=2}^{T'} l_{>0}(\|y_{i,t}-y_{i,t-1}\|_1)(\|\hat{z}_{i,t}-\hat{z}_{i,t-1}\|_2{}^2) + \beta\sum_{i=1}^{N}\sum_{t=2}^{T'} l_{=0}(\|y_{i,t}-y_{i,t-1}\|_1)(\|\hat{z}_{i,t}-\hat{z}_{i,t-1}\|_2{}^2)) \tag{1}$$

$$\hat{z_{i,t}} = z_{i,t}/mod(z_{i,t}) \tag{2}$$

$$\sigma = \begin{cases} e/EPOCH, 1 \leq e \leq EPOCH/2 \\ alog(e/b), EPOCH/2 < e \leq EPOCH \end{cases} \tag{3}$$

Formula (1), (2) and (3) describe our proposed loss function. σ represents the weight of the loss function in different epochs, where $EPOCH$ represents the total number of training epochs and e represents the currently epoch. $l_A(x)$ is an indicator function and $\|\cdot\|_p$ is an L_p norm. α and β are hyperparameters that control the contribution of the additional loss terms, herein we set α as 0.1 and β as 0.03. z and y represent the soft predicted result of the model and the ground truth respectivly. In (3), the specific values of a and b should be obtained according to the boundary conditions: σ should be equal to the limit and slope on the left and right sides at $EPOCH/2$.

In this way, the following two effects can be achieved: 1. Reward the fact that the fast reaction at the event boundary; 2. Penalize the fact that the event should be continuously predicted but not continuous. So that the model constrains the prediction results within the event and the event boundary respectively. Consider that the output of the model is unstable at the beginning of training, and the prediction continuity of the model in events is poor, it may cause a numerical imbalance in the penalty term of wr-TCL, resulting in a serious decline in the overall network training performance. Therefore, we increase the weight σ, so that the weight of wr-TCL tends to 0 at the beginning of training, and increases with epochs and eventually plateaus.

3. EXPERIMENT

3.1. Dataset and feature extraction

All experiments are conducted on the DCASE 2021 domestic environment sound event detection (DESED) dataset, which is composed of real soundscapes and synthesized soundscapes. For real soundscapes, data can be divided into 6 subsets: weakly labeled (1578 clips), unlabeled in domain (14412 clips), synthetic dataset(10000 clips), validation(1168 clips) and evaluation. A strongly labeled dataset from Audioset(3471 clips) is also used for training in the three systems we submitted. The input features used in the proposed system are log-mel pectrograms, which are extracted from the audio signal resampled to 16000 Hz. The log-mel spectrogram is extracted using 2048 STFT windows with a hop size of 256 and 128 Mel-scale filters. As a result, each 10-second sound clip is transformed into a 2D time-frequency representation with a size of (625×128).

Table 1: Final results of models on validation dataset

Model No.	Pre-trained model	Extra dataset	Details	PSDS1	PSDS2
Baseline	1.yes 2.yes 3.no	1.no 2.no 3.yes		0.336 0.351 0.313	0.536 0.552 0.722
Model1	yes	yes	+ FDY CRNN + PANNs predicted framewise feature + wr-TCL + ensemble top5 model	0.481	0.710
Model2	yes	yes	+ FDY CRNN + PANNs predicted framewise feature + wr-TCL + ensemble top10 model	**0.485**	0.725
Model3	yes	yes	+ FDY CRNN + PANNs embedding feature + weak train + ensemble top5 model	0.065	**0.835**
Model4	no	no	+ FDY CRNN + weak train	0.058	0.813

3.2. Experimental settings

The neural networks are trained using the Adam optimizer, with a maximum learning rate of 0.001, and a learning rate rampup during the first 50 epochs. Each model is trained for a total of 200 epochs. In our experiments, we save the best models for PSDS1 and PSDS2 separately, which can be further used for model ensembling. To improve the PSDS2 score, we use the weak train in model3 and model4 [5]. The network archieture of pre-trained model PANNs that we use is ”Cnn14_DecisionLevelMax”[3].

To improve the generalization ability of the model, we perform data augmentation on each model. A total of four data augmentation methods are used: Mixup, Filteraugment, Frame-shift and Specaugment.

3.3. Results

The results for the submitted models on the DESED Real Validation dataset is shown in Table 1. Among the four systems we submit, model 1-3 use the pre-trained model and external data, and model 4 does not use. Model1 and 2 are mainly used to upgrade the PSDS1 score, predicted framewise feature of the PANNs is fused with the feature of FDY-CRNN, and wr-TCL together with MSE and BCE loss constrains the optimization direction of the model. The top 5 models in training are ensembled as model1, and the top 10 models are ensembled as model2. Model 3 concatenat the PANNs embedding features and FDY-CRNN features by weak train to obtain the higher PSDS2 score. Model4 does not use pre-trained model and external data according to the competition requirements, so we only weak trains the FDY-CRNN. In this way, among the four systems we submitted, the highest PSDS1 reached 0.485, and the highest PSDS2 reached 0.835, which is much higher than the three baselines.

4. REFERENCES

[1] N. Turpault, R. Serizel, S. Wisdom, H. Erdogan, J. R. Hershey, E. Fonseca, P. Seetharaman, and J. Salamon, “Sound event detection and separation: A benchmark on desed synthetic soundscapes,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 840–844.

[2] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, “Frequency Dynamic Convolution: Frequency-Adaptive Pattern Recognition for Sound Event Detection,” p. arXiv:2203.15296, 2022.

[3] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, “Panns: Large-scale pretrained audio neural networks for audio pattern recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[4] S. Kothinti and M. Elhilali, “Temporal contrastive-loss for audio event detection,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 326–330.

[5] H. Nam, B.-Y. Ko, G.-T. Lee, S.-H. Kim, W.-H. Jung, S.-M. Choi, and Y.-H. Park, “Heavily augmented sound event detection utilizing weak predictions,” DCASE2021 Challenge, Tech. Rep., June 2021.