THE CAU-ET FOR DCASE 2022 CHALLENGE TECHNICAL REPORTS

Technical Report

Narin Kim, Sumi Lee, Il youp kwak

Chung-Ang University, Department of Applied Statistics,
Seoul, South Korea, {nrgolden, dltnal821,ikwak2}@cau.ac.kr

ABSTRACT

In this technical report, We present a semi-supervised learning method using RCRNN for DCASE 2022 challenge Task 4. We applied three main methods to improve the performance of sound event detection(SED). The first is semi-supervised network using RCRNN based on mean teacher model. The CNN part consists of residual convolution block with a CBAM[1] self-attention module which is stacked 5-layers, and the classification was performed with the RNN part. The second is the application of different data augmentation to features with different types of labels. Mix up, frame shift, time shift, time masking, and filter augmentation were applied to features, mix up was applied differently to the strong label and the weak label, and time masking was applied only to the strong labeled data. The third is to feed features that give different noise to student models and teacher models through data augmentation.The weight of the student model was shared with the teacher model by injecting different feature noise so that it could converge to the global optical faster through consistency loss.

Index Terms— semi-supervised learning, RCRNN, CBAM attention, consistency loss, data augmentation

1. INTRODUCTION

In this technical report, the CAU-ET team describe the SED system for sound event localization and detection task. The main task of this challenge is not only to detect multiple events contained in the corresponding audio data, but also to predict the onset and offset of a event. Strong label with both time stamp and event class information, weak label with only event class, and unlabeled data without any information should be utilized to solve the task, so our team viewed it as a semi-supervised problem like the baseline.

We submit three SED systems: (1) a system stacked with residual convolution block including CBAM attention module in 5-layer which is the CNN part of RCRNN, (2) a system with CBAM attention in the baseline, (3) a system with only data augmentation added to the baseline. The submitted system is based on a semi-supervised model which is Mean Teacher[2] model (beasline). We try to improve the generalization performance of the model by applying time shift, frame shift, mixup, time masking, and filter augmentation to inject noise to the features differently. In addition, features to which filter augmentation was applied differently feed to network as input and the network trained through consistency loss. Among the submitted systems, the system with the residual convolution block had the best performance on psds scanario2, and the system with CBAM attention is the highest psds scanario1. they are better than baseline performance.

2. METHOD

2.1. Residual Convolution Block with CBAM attention

Since the task of this challenge requires to detect the sound events in each audio clip and predicting the onset and offset of the event(localization) , we used the CBAM self-attention module to learn important parts of the event more attentively. It trys to learn weights by focusing on more important parts in log mel spectrogram to detect events onset and offset. In the first paper of resnet[3], skip connections are made before convolution output and ReLU functions in the residual convolution block, but experiments after the initial model show that full pre-activation, which is a skip connections after two calculations in batch normalization, ReLU, and convolution order. Therefore, we used full pre-activation when building a residual convolution block. As shown in Figure 2, the cbam attention module was added to perform self-attention after the second operation and then the information was aggregated through average pooling layer to extract a feature map for log-mel energy. The overall architecture of the RCRNN we used is shown in Figure 1.

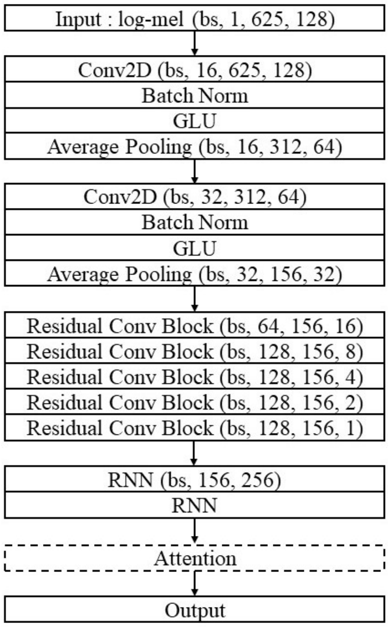
Input : log-mel (bs, 1, 625, 128)
Conv2D (bs, 16, 625, 128)
Batch Norm
GLU
Average Pooling (bs, 16, 312, 64)
Conv2D (bs, 32, 312, 64)
Batch Norm
GLU
Average Pooling (bs, 32, 156, 32)
Residual Conv Block (bs, 64, 156, 16)
Residual Conv Block (bs, 128, 156, 8)
Residual Conv Block (bs, 128, 156, 4)
Residual Conv Block (bs, 128, 156, 2)
Residual Conv Block (bs, 128, 156, 1)
RNN (bs, 156, 256)
RNN
Attention
Output

Figure 1: RCRNN architecture

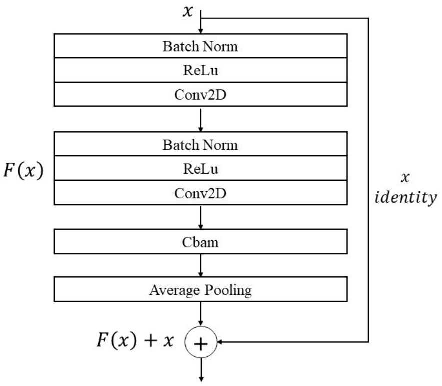

Figure 2: Residual Convolution Block architecture

2.2. Data Augmentaion

To enhance the performance of the model, various augmentation methods were applied and experimented as follows:

Filter Augmentation[4]: Filter augmentation is an augmentation method that takes into account various acoustic conditions (such as different microphones and speakers, and the relative position between microphones and speakers). It mimics the acoustic conditions in which a human can classify sounds well using filters. There are two types of Filter Augmentation, and we applied the 'step' type method that divides the features at equal intervals and applies different filters to each. The algorithm randomly selects a frequency boundary number n, and then divides the Mel-spectrogram into n+1 frequency bands. A factor to be multiplied by the amplitude of the Mel spectogram corresponding to each determined frequency band is randomly selected. The number of frequency bands and the dB range for the random factor are hyperparameters. We applied filter augmentation differently to the student model and teacher model of the mean teacher.

time mask / frame shift / time shift : In the case of time mask, values for the position where masking is applied to the time axis and the width of the mask were uniformly extracted and masked. Frame shift and time shift were also randomly extracted with boundary values to which data shifting will be applied. These augmentation were applied equally to weak data and strong data.

Mixup[5] : Mixup can improve the performance of deep learning in many 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 is formulated as follows:

$$\widetilde{X} = \lambda X_i + (1-\lambda)X_j \tag{1}$$

$$\widetilde{Y} = \lambda Y_i + (1-\lambda)Y_j \tag{2}$$

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. We applied mixup differently for strong data and Weak data, unlike previous augmentation methods. The experiment was conducted by changing the parameter value that determines the ratio of the data to which the mixup is applied.

Table 1: Experimental results with data augmentation

	Augmentation	PSDS1	PSDS2
model1	mixup rate = 0.5 FilterAug:dB = - 4.5 ∼ 6 #band = 2 ∼ 5	0.372	0.592
model2	mixup rate = 0.5 FilterAug:dB = - 4.5 ∼ 6 #band = 2 ∼ 5	0.377	0.585
model3	mixup rate = 0.8 FilterAug:dB = - 7.5 ∼ 6 #band = 2 ∼ 3	0.373	0.571

2.3. feature noise with consistency loss

Semi-supervised learning is performed by feeding features to the student model and the teacher model that have different noise levels through filter augmentation. The supervised loss is binary cross entropy and semi-supervised loss is consistency loss (MSE) to optimize both at 1-stage as shown in Equation 3.

$$loss_{Total} = loss_{supervised:BCE} + loss_{semi\text{-}supervised:MSE} \tag{3}$$

3. EXPERIMENTS

3.1. Data

The training set consists of (1) synthetic strongly labeled data (10000), (2) unlabeled in domain data (14412), (3) weakly labeled (1420) and (4) strong real data (3470). The subset of weak data was used for training and remainder is used for validation. Train and validation were split by 0.9. Strong real data is external data, and there are two systems which include it to train the network and other except for it. The validation set is (1) synthetic strongly labeled (2500), (2) weakly labeled (158), and the performance of the model was evaluated by strongly labeled.

- train set
  1. synthetic strongly labeled data (10000)
  2. unlabeled in domain data (14412)
  3. weakly labeled (1420)
  4. strong real data (3470)
- validation set
  1. synthetic strongly labeled (2500)
  2. weakly labeled (158)
- test set
  1. strongly labeled (1168)

3.2. feature extraction

The log-melspectrogram produced in 128 dimensions of the mel bin size extracted with the torchaudio module is a feature of the mean -teacher model. Log-melspectrogram features were extracted with a sample rate of 16000. The input segment is 10 seconds long,

3.3. training

The hardware used for training was 1 RTX 2080 Ti, and the Adam optimizer was used as the optimizer. All models were experimented with up to 200 epochs and early stopping was applied based on the intersection F1 score.

4. RESULTS

The CAU-ET submitted three SED systems for DCASE 2022 Challenge Task 4. As shown in Table2, the PSDS scenario1 performance of the model with the CBAM attention module applied to each layer of the CNN of the baseline CRNN was the best with 0.377. In PSDS scenario 2, RCRNN using residual convolution block was the best performance with 0.592.

Table 2: Final results of experiments

	PSDS1	PSDS2
model1	0.372	0.592
model2	0.377	0.585
model3	0.373	0.571

5. ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIP) (No. NRF-2020R1C1C1A01013020 and NRF-2017R1A2B2007216).

6. REFERENCES

[1] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” 2018. [Online]. Available: https://arxiv.org/abs/1807.06521

[2] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” 2017. [Online]. Available: https://arxiv.org/abs/1703.01780

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” 2016. [Online]. Available: https://arxiv.org/abs/1603.05027

[4] 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.

[5] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” 2017. [Online]. Available: https://arxiv.org/abs/1710.09412