SOUND EVENT DETECTION SYSTEM USING FIXMATCH FOR DCASE 2022 CHALLENGE TASK 4

Technical Report

Changmin Kim

LG Electronics
Seoul, South Korea
changmin0.kim@lge.com

Siyoung Yang

LG Electronics
Seoul, South Korea
siyoung.yang@lge.com

ABSTRACT

This technical report proposes a sound event detection (SED) system in domestic environments for DCASE 2022 challenge task 4. In this system, the training method consists of two stages. In the stage 1, mean teacher (MT) and interpolation consistency training (ICT) are used. In the stage 2, FixMatch is additionally applied. We adopted the frequency dynamic convolution recurrent neural network (FDY-CRNN) structure as our model. In order to further improve the performance of polyphonic sound detection score (PSDS) scenario 2, three techniques were used. First, we applied a temperature parameter to the sigmoid function to obtain soft confidence value. Second, we used a weak SED that is a method that uses only weak predictions and sets the timestamp equal to the total duration of the audio clip. Third, the FSD50K dataset was added to the weakly labeled dataset, which helped the PSDS scenario 2. As a result, we obtained the best PSDS scenario 1 of 0.473, and best PSDS scenario 2 of 0.695 on the domestic environment SED real validation dataset.

Index Terms— Polyphonic Sound Event Detection, Semi-Supervised Learning, FixMatch

1. INTRODUCTION

Sound event detection (SED) is to find out if an audio clip has a sound of interest, and when the sound starts and ends. The development of deep learning has made it possible to solve various problems. Recently, the SED system has also been improved a lot through deep learning. Although deep learning requires large amount of data, it is difficult to obtain clearly labeled data for audio data compared to other vision data or text data. Therefore, semi-supervised learning, which learns using both unlabeled and labeled data, is also attracting attention. Detection and classification of acoustic scenes and events (DCASE) challenge task 4 addresses the problem of SEDs with little or no labeled data. Find 10 different sounds in this 10-second audio clip: alarm/bell/ringing sounds, blender, cat, plate, dog, electric razor/toothbrush, splashing water, running water, horse, vacuum cleaner and more. The dataset provided for training includes not only strong label data including the presence or absence of a target sound, but also weak label data providing only the presence or absence of a target sound and unlabeled data without any information. Up to now, the task 4 has mainly used ordinary teachers and performed well in interpolation consistency training (ICT) [1] and self-study. FixMatch is a kind of semi-supervised learning method proposed by Sohn [2]. Although it shows better performance than the existing average semi-supervised learning method, FixMatch has been proposed only for classification problems and has not been used for detection problems. Therefore, in this report, we propose a modified FixMatch method for the detection problem. As models used for SED, methods such as selective kernel network (SKNet) [3] and residual convolution recurrent neural network (RCRNN) [4] have been proposed, starting with the baseline CRNN. In our system, we applied the frequency dynamic convolution recurrent neural network (FDY-CRNN) [5] proposed by Nam recently, because it outperforms other models. In addition, we changed the sigmoid function to use weak prediction and used weak SED and added weakly labeled data using FSD50K [6] in attempt to improve polyphonic sound detection score (PSDS) scenario 2 performance.

2. PROPOSED METHOD

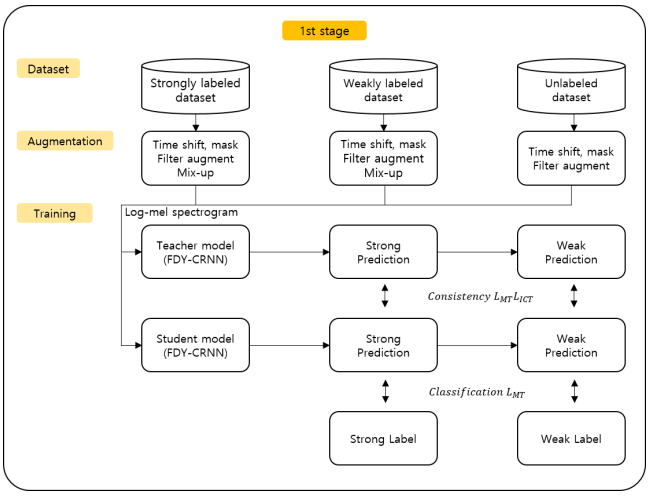

Figure 1: 1st stage of the proposed method

We propose a method of learning in two stages. In the 1st stage, data augmentation is applied to the strongly labeled dataset, weakly labeled dataset, and unlabeled dataset. At this point, the mix-up is applied, except for the unlabeled dataset [7]. Then log-mel spectrogram feature extraction is performed on the augmented dataset. When extracting features, we used 2048 sample frame length, 256

sample hop length, and 128 mel-frequency bands. Both the teacher model and the student model were implemented with FDY-CRNN. The initialized weight is different and the dropout is set to 0.5, so even if the same feature is input, a different output will come out. Consistency loss is obtained by comparing the values passed through the teacher model and the student model for the same data, respectively. In this case, both ICT and MT are used. The output of the student model is compared with the actual label to obtain a classification loss. The model is trained in the direction of decreasing the values of the consistency loss and the classification loss.

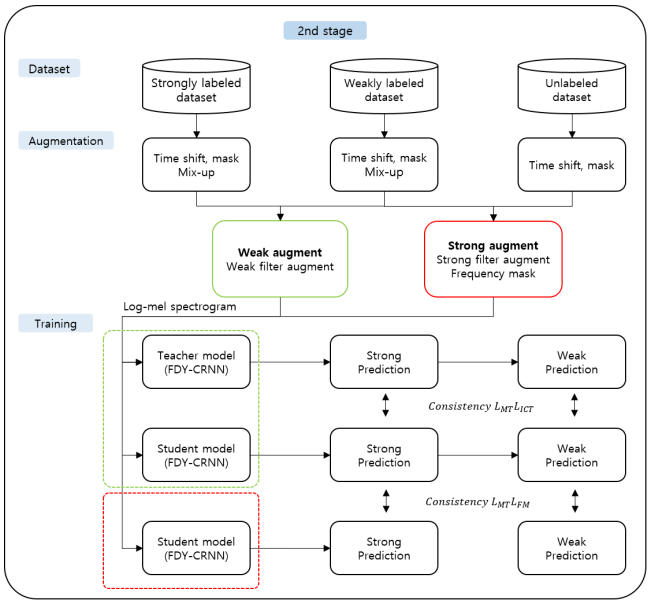

Figure 2: 2nd stage of the proposed method

The 2nd stage is almost same to the 1st stage, except that weak and strong augmentations are applied separately when augmenting the data. Details of augmentation will be described later in 2.3. As in the 1st stage, weak augment feature is used to obtain consistency loss through MT and ICT methods. The strong augment feature is compared with the prediction of the weak augment feature, and the loss value is calculated using the FixMatch method.

2.1. FDY-CRNN

We adopted the FDY-CRNN structure as our model. This model consists of 7 CNN-based layers and 2 BiGRU layers, like the CRNN structure in baseline. However, CNN-based layers are different from baseline models. Unlike the baseline model, The CNN-based layer of FDY-CRNN applies a kernel that adapts to each frequency bin of the input to remove transform invariance of 2D convolutions along the frequency axis.

2.2. ICT

Since ICT also uses a student model and teacher model similar to MT, it would be better to explain the MT method in detail first. The MT calculates binary cross entropy (BCE) for labeled data and the mean squared error (MSE) between each outcome in unlabeled data via a student model and a teacher model. As in (1), the MT loss is obtained by adding the BCE loss and the MSE loss. ICT uses the mix function (2) to calculate the loss of unlabeled data. The final loss of ICT is calculated as (3).

$$L_{MT} = \sum_{x \in B} BCE(f_\theta(x), label) + \sum_{x \in B} MSE(f_\theta(x), f_{\theta'}(x)) \tag{1}$$

L_{MT} indicates MT loss. f_θ is the student model and $f_{\theta'}$ is the teacher model.

$$Mix_\lambda(a, b) = \lambda * a + (1 - \lambda) * b \tag{2}$$

The Mix function mixes a and b according to the λ ratio.

$$L_{ICT} = \sum_{x_1, x_2 \in B} MSE(f_\theta(Mix_\lambda(x_1, x_2)), Mix_\lambda(f_{\theta'}(x_1), f_{\theta'}(x_2))) \tag{3}$$

L_{ICT} indicates ICT loss.x_1 and x_2 are two samples randomly selected in the batch B.

2.3. FixMatch

FixMatch is described in Figure 3. So far, FixMatch has never been applied to SEDs due to some issues. Therefore, we found three solutions to use FixMatch as follows.

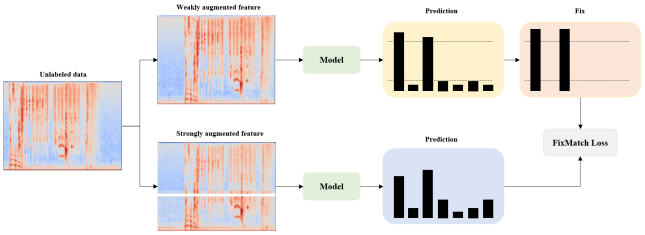

Figure 3: Strongly augmented features and weakly augmented features were passed to the same model. If all confidence values of the weakly augmented feature is above threshold, pseudo labeling is performed. BCE is calculated for prediction values of pesudo labels and strongly augmented features, and added to the total loss.

2.3.1. Augmentation

To apply FixMatch, strong and weak augmentation are required. Since FixMatch was originally proposed for image data, augmentation was required for audio. In this report, time shifting, time masking, and mix-up were applied in common. Weak filter augment was used as weak augmentation. For strong augmentation, frequency masking and strong filter augment were used. Filter augment is an augmentation method proposed by Nam [8], which changes the signal strength for each frequency. Weak and strong augmentation changed 2dB and 6dB, respectively. Frequency masking randomly masked 16 out of 128 mel bins.

2.3.2. Weakly labeled data

In addition to strongly labeled data and unlabeled data, the task 4 has weakly labeled data. Weak prediction of weakly labeled data calculated prediction loss, and strong prediction calculated consistency loss.

2.3.3. Threshold

FixMatch had different conditions in image classification and the SED in task 4.In the case of image classification, the model that classify 10 images always had a target image which is one of the 10 images and didn't have more than one images. In the case of the SED, However, non-target sounds may appear and more than one sound may appear at the same time. We modified FixMatch for the SED instead of image classification. We used two thresholds for in Figure 3.

$$L_{FM} = \sum_{x \in B} BCE(f(x_{strongAug}), f(x_{weaklyAug})) \quad (4)$$

L_{FM} indicates FixMatch loss function.f indicates student model. $x_{strongAug}$ indicates a strongly augmented sample and $x_{weaklyAug}$ indicates a weakly augmented sample in batch B.

$$TotalLoss_{stage_1} = L_{MT} + L_{ICT} \quad (5)$$

$$TotalLoss_{stage_2} = L_{MT} + L_{ICT} + L_{FM} \quad (6)$$

In the baseline model applying the modified FixMath with three techniques, PSDS1 increased from 035 to 0.39. It had higher PSDS1 than the baseline model with ICT. Performance was better when using both ICT and FixMatch. Specifically, it proceeds in two steps. In the first step, the model is first trained with ICT to make the confidence values stable. The total loss value at this time is expressed as (5). The second stage uses FixMatch with ICT like (6). As such, the performance was improved compared to using ICT only or FixMatch alone.

2.4. Techniques to improve PSDS2

2.4.1. Temperature parameter

We adopted the temperature parameter [9] that was added to the last layer of the model, the sigmoid. The temperature parameter smooths the result of the sigmoid function and the extreme confidence value. In our system, a value of 10 showed the highest performance as a temperature parameter as a result of the experiment. It is expressed as (7).

$$y_i = sigmoid(z_i/10) = \frac{1}{1 + exp(-z_i/10)} \quad (7)$$

where z_i, y_i means the confidence and smoothed detection output for the event class. i.

2.4.2. Weak SED

We also used the mild SED used in last year [10]. Weak SED is a method that uses only weak predictions and sets the timestamp equal to the total duration of the audio clip. That is, if the weak prediction corresponding to a label exceeds the threshold, it predicts that the label will be present from beginning to end of the audio clip. This method significantly lowers PSDS1 but raises PSDS2.

2.4.3. FSD50K dataset

Finally, the weakly labeled data of FSD50K was added [6]. The FSD50K dataset had a lot of weakly labeled data before it was merged with the background data, and when combined with the existing data, PSDS1 decreased but PSDS2 increased.

3. EXPERIMENTAL RESULTS

3.1. Model training

We used xavier initialization. For optimization, ADAM was used. Dropout used 0.5. As explained in Chapter 2, the learning process consists of stage 1 using ICT only and stage 2 using ICT and FixMatch together. In Stage 1, the learning rate was increased exponentially to 0.001 for 50 epochs and maintained at 0.001 for 150 epochs. In Stage 2, the learning rate fluctuated between 0.00001 and 0.001 using a cosine annealing scheduler for 300 epochs.

3.2. Discussion

As shown in Table 1, the performance of the FDY-CRNN models significantly increased compared to the baseline model. When ICT was added as a semi-supervised learning method, the performance increased compared to when only MT was used. Finally, adding FixMatch improved the performance a bit.

Model	Semi-supervised learning	PSDS1	PSDS2
Baseline	MT	0.373	0.549
FDY-CRNN	MT	0.444	0.656
FDY-CRNN	MT + ICT	0.455	0.666
FDY-CRNN	MT + ICT + FixMatch	0.467	0.665

Table 1: Performance according to the application of FDY-CRNN, ICT, and FixMatch

Temperature	Weak SED	FSD50K	PSDS1	PSDS2
O	X	X	0.410	0.725
X	O	X	0.062	0.781
X	X	O	0.445	0.684
O	O	O	0.069	0.809

Table 2: Performance according to application of three techniques to improve PSDS2

Experimental results for three techniques to improve PSDS2 are presented in Table 2. When the temperature parameter was applied, the performance of PSDS1 decreased slightly, but the performance of PSDS2 increased. PSDS2 also increased when weak SED was applied. However, PSDS1 performance was significantly reduced. PSDS2 slightly increased when FSD50K was applied. And PSDS1 also decreased slightly.

4. CONCLUSION

FixMatch was applied in a semi-supervised learning method. FDY-CRNN was applied to the model. We also applied three techniques to improve PSDS2 performance.

Submission	Ensemble	PSDS2 techniques	PSDS1	PSDS2
Kim_LGE_1	9	X	0.473	0.693
Kim_LGE_2	10	X	0.473	0.695
Kim_LGE_3	23	O	0.068	0.830
Kim_LGE_4	23	Temperature only	0.354	0.756

Table 3: Performance and number of ensembles used in four submissions

Our submissions are shown in Table 3. Two models with good PSDS1 performance and two models with good PSDS2 performance are submitted. When ensemble was performed, We used various models: A model with a different FixMatch threshold, a model with a modified FixMatch loss weight, or a model without FixMatch. When various models with these parameters adjusted were used together, the performance was better than when using a single model.

5. REFERENCES

[1] C. Y. Koh, Y. S. Chen, Y. W. Liu, and M. R. Bai, "Sound event detection by consistency training and pseudo-labeling with feature-pyramid convolutional recurrent neural networks," in *Proc. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 376–380.

[2] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C. L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, vol. 33, pp. 596–608, 2020.

[3] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 510–519.

[4] N. K. Kim and H. K. Kim, "Polyphonic sound event detection based on residual convolutional recurrent neural network with semi-supervised loss function," *in IEEE Access*, vol. 9, pp. 7564–7575, 2021.

[5] H. Nam, S. H. Kim, B. Y. Ko, and Y. H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," *arXiv preprint arXiv:2203.15296*, 2022.

[6] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *in IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

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

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

[9] X. Zheng, H. Chen, and Y. Song, "Zheng ustc teams submission for dcase2021 task4 semi-supervised sound event detection," DCASE2021 Challenge, Tech. Rep., 2021.

[10] 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," *arXiv preprint arXiv:2107.03649*, 2021.