IMPROVED PSEUDO-LABELING METHOD FOR SEMI-SUPERVISED SOUND EVENT DETECTION

Technical Report

Yaguang Gong, Changlong Li, Xintian Wang, Lu Ma, Song Yang, Zhongqin Wu,

TAL Education Group, China

ABSTRACT

This report illustrates a framework for the DCASE2021 task4 - Sound Event Detection. The proposed framework is built on the pseudo-labeling method widely applied for semi-supervised learning(SSL) tasks. The proposed method synthesizes weak pseudo-labels for the large amount of unlabeled data by utilizing the model's predictions on weakly augmented spectrograms. Weak pseudo-labels are then used as supervision for strongly augmented spectrograms of the same sample. Along to this main contribution, this work introduces data augmentation techniques including random frequency masking and time shifting, training techniques such as class-specific weighted loss, and model ensemble techniques. Experimental results demonstrate that the proposed method achieves PSDS of 0.407/0.653(scenario1/scenario2) on the validation set, which presents superior performance against the baseline score 0.342/0.527.

Index Terms— Sound event detection, Semi-supervised learning, Pseudo-label method

1. INTRODUCTION

The technical report describes solution systems for DCASE2021 Challenge Task4: sound event detection(SED) and separation in domestic environments. The target of this task is to train a model capable of detecting specific domestic sound events with weakly labeled, unlabeled real-world data and synthetic data which has strong labels. Our submission systems basically follow the baseline architecture[1], while considering the characteristics of evaluation metrics, we adopt several techniques to improve the performance:

- improved pseudo-labeling methods;
- data augmentation techniques, for instance specAug[2] and time shift[3];
- class-specific weighted loss;
- post-processing refinement;
- model ensemble.

We carry out evaluation experiments on DCASE2021 task4 validation set to verify the effectiveness of those mentioned above. The result demonstrates that our submission greatly outperform the baseline system, achieving the best PSDS-scenario1 of 0.407 and PSDS-scenario2 of 0.653.

2. PROPOSED METHOD

The CRNN[4] architecture has previously achieved great performance on SED tasks. With the effective feature extraction ability of CNN and strong time-dependency modeling ability of RNN, CRNN combines both of them to get a higher promotion which is why we choose it as our basic system. Our network is mainly the same as that of baseline system.

2.1. Data preparation

We down-sample the original audio to 16kHz and generate 128-dimensional log-Mel filter-banks. The window size and hop size are 2048 and 256 respectively. All the training audio are aligned to 625 frames which corresponds to 10 seconds. Finally, features are normalized along the whole training set before sent into the network as input. For the sake of robustness, we diversify training data by utilizing several data augmentation methods such as mixup[5], time shift and specAug.

2.2. Semi-supervised method

Due to the large mount of unlabeled data, semi-supervised strategy is essential. Mean-teacher[6] is widely applied in previous systems and achieve obvious effect which is also included in our systems. In addition, we introduce an improved pseudo-labeling method enlightened by Fixmatch[7] of image classification task. In short, the total loss function consists of two parts, a supervised loss applied to labeled data and an unsupervised loss. The latter combines pseudo-labeling and consistency regularization in the meantime. Weakly-augmented spectrograms are fed in the model to generate pseudo labels against which strongly-augmented version predictions are used to enforce entropy loss.

2.3. Training techniques

As we analyzed, the PSDS[8] metrics of two scenarios both suffer greatly from instability across classes. So we attempt to assign diversified weights for the loss of each class, which eventually boost the performance a little. Also, we find that post median filtering would in certain extent harm the performance mainly due to the nature of metrics that continuity of the frames output doesn't matter.

2.4. Model ensemble

To utilize advantages of different models, we select several top systems and average the raw outputs before post-processing. Every model is trained with different training settings in order to fuse unique generalizing ability.

3. EXPERIMENTAL EVALUATION

3.1. Experimental conditions

We conduct experimental evaluations on DCASE2021 Task4 dataset[9]. The dataset contains 1578 audio clips with weak label, 10000 synthesized audio clips using baseline script and 14412 unlabeled audio clips. We choose AdamW optimizer with learning rate of 0.001. The model is training for 36k steps and the beginning 2.5k steps are for warming up[10]. For mean-teacher, the batch size is 128 consisting of 32 synthesized, 32 weakly labeled and 64 unlabeled data. For pseudo labeling, unlabeled data need to be of a higher percentage, which we put 32, 32, 128, respectively in a total batch of 192.

3.2. Experimental results

3.2.1. Effects of SSL methods

First, we investigate the merits of different semi-supervised learning strategies. By keeping all the same with baseline except replacing mean-teacher with improved pseudo labeling, the metrics rise higher under both scenarios as Table 1 shows.

Method	PSDS scenario1	PSDS scenario2
baseline	0.342	0.527
i-pseudo-labeling	0.360	0.553

Table 1: Effects of SSL methods

3.2.2. Effects of data augmentation

Next, we investigate the effects of data augmentation techniques . On the basis of baseline system, specAug and time shift are respectively enforced. Table 2 shows the improvement.

Method	PSDS scenario1	PSDS scenario2
baseline	0.342	0.527
+specAug	0.358	0.550
+time shift	0.358	0.563
+specAug+time shift	0.368	0.576

Table 2: Effects of data augmentation

3.2.3. Effects of post-processing strategies

Next, we attempt to study the effects of post-processing. Using improved pseudo-labeling model, with respect to post median filter, we consider 3 situations: global length of baseline, class-wise length counted from training set and without post-processing. It turns out that filtering may even hurt the system performance as Table 3 shows.

3.2.4. Effects of model ensemble

Finally, we attempt to study the effects of model ensemble. Table 4 shows the results of best single mean-teacher model, best single improved pseudo-label model and best ensemble model. The contribution of ensemble is quite obvious.

Method	PSDS scenario1	PSDS scenario2
global length	0.367	0.599
class-wise length	0.381	0.614
without filtering	0.394	0.624

Table 3: Effects of post-processing strategiese

Method	PSDS scenario1	PSDS scenario2
mean-teacher	0.382	0.619
i-pseudo-labeling	0.394	0.624
model ensemble	0.407	0.653

Table 4: Effects of model ensemble

4. CONCLUSION

In this report, we propose and describe a submission system for DCASE2021 task4. The system involves an improved pseudo-labeling method which also combines regularization consistency. Also, training techniques such as class-specific weighted loss and post-processing filtering strategy are introduced in terms of the nature of evaluation metrics. Extensive experiments and ablation study demonstrate the feasibility of our methods.

5. REFERENCES

[1] S. Wisdom, H. Erdogan, D. P. W. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the fuss about free universal sound separation data?" in *in preparation*, 2020.

[2] 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," 2019.

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

[4] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio Speech Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

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

[6] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," 2017.

[7] K. Sohn, D. Berthelot, C. L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," 2020.

[8] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, "A framework for the robust evaluation of sound event detection," *arXiv preprint arXiv:1910.08440*, 2019. [Online]. Available: https://arxiv.org/abs/1910.08440

[9] N. Turpault, R. Serizel, A. Parag 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*, New York City, United States, October 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[10] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with restarts," 2016.