CHT+NSYSU SOUND EVENT DETECTION SYSTEM WITH DIFFERENT KINDS OF PRETRAINED MODELS FOR DCASE 2022 TASK 4

Technical Report

Sung-Jen Huang[1], *Chia-Chuan Liu*[1], *Chia-Ping Chen*[1], *Chung-Li Lu*[2], *Bo-Cheng Chan*[2], *Yu-Han Cheng*[2], *Hsiang-Feng Chuang*[2]

[1] National Sun Yat-Sen University, Taiwan,
{m093040011,m103040063}@student.nsysu.edu.tw,
cpchen@cse.nsysu.edu.tw
[2] Chunghwa Telecom Laboratories, Taiwan,
{chungli,cbc,henacheng, gotop}@cht.com.tw

ABSTRACT

In this technical report, we describe our submission system for DCASE 2022 Task4: sound event detection in domestic environments. We proposed two kinds of systems. One is trained by combining the mean teacher framework and knowledge distillation (one student model and two teacher models) without external data. While training this system, we first trained a mean teacher model to be a pretrained model. Our next step is to select the better one, the teacher or student model, to be the trained model for knowledge distillation. Afterword, we trained another mean teacher model with a different architecture using knowledge distillation. Finally, we repeat the select model step and knowledge distillation several times. The mean teacher model in the final round is composed of a VGG block, selective kernels and a clip level consistency branch. Comparing to the PSDS-scenario1 of 35.1% and PSDS-scenario2 of 55.2% of the baseline system trained without external data, the ensemble of this kind of system can achieve 43.7% and 68.0%, respectively.

The other system can be separated into two parts. The first part is the top three layers of pretrained PANNs, while the second part is a similar system to baseline with only three convolution blocks. Then we trained the whole system (included PANNs) with DESED data. Ensembleing this system, the PSDS-scenario1 and 2 of 46.5% and 76.7% outperforms the baseline system (trained with AST embedding) of 31.3% and 72.2%.

Index Terms— Sound event detection, pretrained model, mean teacher, knowledge distillation, PANNs

1. INTRODUCTION

This technical report describes the submission system for DCASE2022 Challenge Task4. In Task4, the goal is to detect a sound event in a domestic area. The challenge allowed participants to use external data this year, but each team needs to submit at least one system trained without external data. Following this rule, we submitted a ensemble system incorporating the mean teacher method and knowledge distillation [1] for the required system trained without external data. To train this system, we first trained a mean teacher model. In the second step, we choose the better model between the student and teacher to use for knowledge distillation. We utilized knowledge distillation to train a new mean teacher model based on a different architecture in the third step. In the last step, we repeated the second and third step several times. We can improve the ability of the model without increasing inference time in this way.

On the other hand, we developed another kind of system combining part of PANNs (Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition) [2] and part of the baseline CRNN system. Structure of this system can be summarized as follows: the top three convolution blocks of PANNs with their pretrained parameters, three convolution blocks as in the baseline system and the same bi-directional gated recurrent units. We submit a single model system and two ensemble systems in this kind of network. The two ensemble systems have the highest PSDS-scanerio 1 and 2 in our experiment, respectively.

In the rest of this report, section 2 describes the dataset and signal processing. The following section explains more details and the score on validation set of our submission. The conclusion and discussion is in section 4.

2. DATASET AND SIGNAL PROCESSING

2.1. Dataset

Domestic Environment Sound Event Detection Dataset (DESED dataset) is an open dataset that we use in this work. The dataset contains data with strong labels, weak labels, and unlabeled data. A strong label for an acoustic clip provides the class of each sound event within the clip, along with the corresponding time stamps (start and end times). A weak label provides only the classes of the sound events, i.e. without any time stamps. The unlabeled data segment consists of acoustic clips without any labels. There are ten classes of domestic environment sound events: Alarm/bell/ringing, blender, cat, dishes, dogs, electric shaver/toothbrush, frying, running water, speech and vacuum cleaner. The data set used for system training consists of 10,000 synthetic and 3470 real samples (sound clips) with strong labels, 1,578 sound clips with weak labels, and 14,412 unlabeled samples.

2.2. Signal processing

For the system incorporating mean teacher and knowledge distillation, the signal processing is the same as baseline, except that we

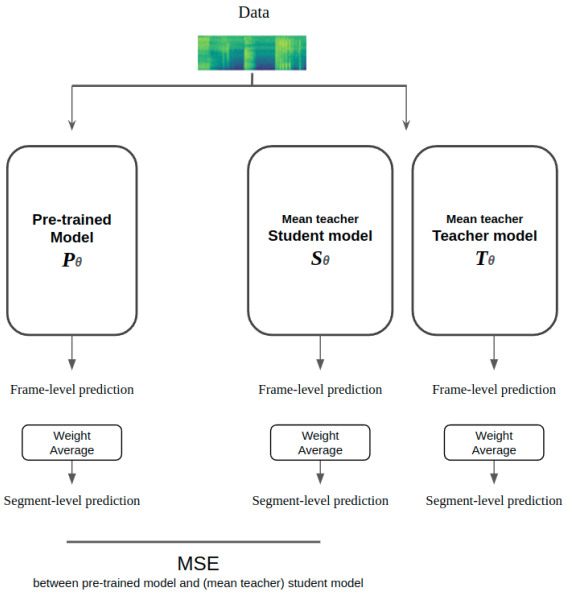

Figure 1: Incorporating mean teacher model and knowledge distillation

changed the padding length from 10 seconds to 12. And for the system combining PANNs and the baseline CRNN, we first resample the 44.1k audio clips to 32k. Then we generate the log mel-spectrogram with the same setting in PANNs.

3. PROPOSED METHODS

3.1. System using knowledge distillation

Knowledge distillation is frequently used to compress a model. It often exploits a pre-trained complex model with a better score to teach a simple model with a lower score. The two models are often referred to as the "teacher" and "student", which is in line with the mean teacher framework. Similarly, there is a consistency loss to make the student model simulate the teacher model as well. In this way, we could transfer knowledge from teacher to student. The main difference between knowledge distillation and the mean teacher method is whether the teacher model will tune its parameters or not. Unlike the mean teacher method, the teacher model in knowledge distillation is pre-trained and will not tune its parameters in the training.

In this paper, we use knowledge distillation in a slightly different way. Specifically, the size of the pre-trained model to be learned from is not necessarily bigger than the model learning from the pre-trained model. Furthermore, we combine the mean teacher and knowledge distillation methods in training, so there are 2 teacher models and one student model in total. The structure of this model is in figure 1.

For this submission system, we first trained a mean teacher model (VGGSK) composed of a VGG block [3] and four residual blocks with selective kernels (SK) [4]. Then we used the teacher model of this mean teacher model to train another mean teacher model composed of 5 RepVGG blocks [5] through knowledge distillation. That is, there is a student and a teacher model of mean teacher RepVGG model and another VGGSK teacher model used for knowledge distillation. The way to incorporate these two methods is pretty simple. We just added a term calculating the mean square error between the student model (RepVGG) and the pre-trained model (VGGSK) to the previous consistency loss between the student and teacher of mean teacher model.

	training method	PSDS1	PSDS2
VGGSKCCT	MT	0.390	0.638
VGGSKCCT	MT+KD	0.404	0.652
VGGSKCCT	MT+KD+SCT	0.424	0.674
VGGSKCCT-ensemble	average	0.437	0.680

Table 1: PSDS of submission system without external data

	model count	PSDS1	PSDS2
PACRNN	1	0.451	0.734
PACRNN	4	**0.465**	0.760
PACRNN	5	0.457	**0.767**

Table 2: PSDS of submission system combining PANNs and CRNN

$$L_{KD} = MSE[\theta_w(X), \theta_w^{KD}(X)] \ + \ MSE[\theta_s(X), \theta_s^{KD}(X)] \tag{1}$$

where θ denotes prediction of student model and θ^{KD} denotes the prediction of the pretrained model, and the subscript w and s is weak prediction and strong prediction, respectively. The consistency loss of the model will be like (2).

$$L_{consistency} = w \ * \{L_{consistency} + \ L_{KD}\} \tag{2}$$

After training the mean teacher model of RepVGG described above, we trained another mean teacher model in the same way. In this round, the mean teacher model is a structure adding a clip level consistency branch [6] to VGGSK, and the pretrained model is RepVGG. To further improve the score, we also used frame shifting in shift consistency training [7] in this round. Note that the whole training did not use any external data, and every round we applied the mixup method for data augmentation as in the baseline system.

For model ensembling, we simply average the prediction setting temperature parameters (to 2) which is used in the system of [8]. Five training checkpoints without SCT and one using SCT were ensembleed. Table 1 shows the PSDS of the basic mean teacher model of (VGGSK adding clip level consistency branch), the single model using the methods described above, and the ensemble system based on 6 checkpoints on the validation set. We submitted this ensemble system as the one without external data.

3.2. System using pretrained PANNs

The system combining pretrained PANNs and baseline CRNN systems consists of the top three convolution blocks of PANNs with their pretrained parameters, three 3×3 convolution blocks with 256, 128, and 128 channels, pooling sizes of 1×2, and RNN layers identical to baseline. The architecture can be refer to figure 2. Note that only PANNs are pretrained, and this part will be finetuned while training. Table 2 shows the result on the validation set of single model and ensemble system. We average different checkpoints to get two ensemble systems with the highest PSDS1 and 2, respectively.

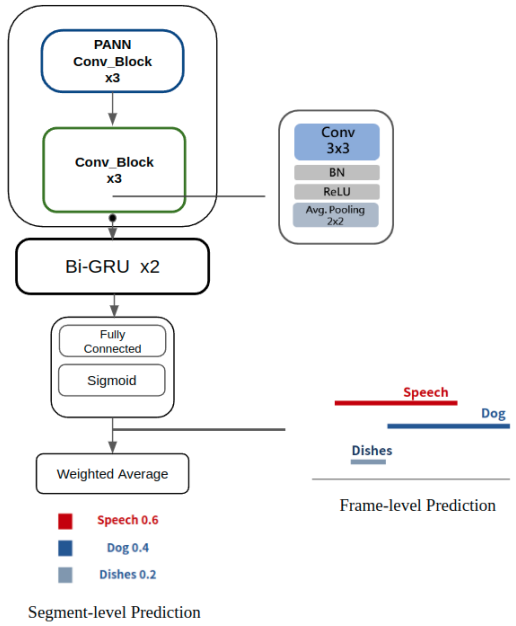

Figure 2: PACRNN: combination of PANNs and baseline CRNN

4. CONCLUSION & DISCUSSION

In our experiment, we found that using a pretrained model with or without external data can both increase model ability. It will significantly improve the model while using a pretrained model with external data. But data augmentation will not have much effect on the system based on a pretrained model which is already trained on lots of external data. Besides, averaging the prediction of different checkpoints can ensemble the model and improve the model's robustness to get a better score.

While generating the output of the evaluation set, we found that the prediction is wrong for five minutes sound clips due to the problem with the code. The max padding length setting in the code while decoding the prediction from frames to seconds limits the transformation to be less than padding length (10/12 seconds in baseline/our setting). To obtain the true prediction, we canceled this limitation.

5. REFERENCES

[1] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[2] 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," *CoRR*, vol. abs/1912.10211, 2019. [Online]. Available: http://arxiv.org/abs/1912.10211

[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," *CoRR*, vol. abs/1903.06586, 2019. [Online]. Available: http://arxiv.org/abs/1903.06586

[5] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," *CoRR*, vol. abs/2101.03697, 2021. [Online]. Available: https://arxiv.org/abs/2101.03697

[6] L. Yang, J. Hao, Z. Hou, and W. Peng, "Two-stage domain adaptation for sound event detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 230–234.

[7] 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 *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 376–380.

[8] X. Zheng, H. Chen, and Y. Song, "Zheng ustc team's submission for dcase2021 task4 – semi-supervised sound event detection," DCASE2021 Challenge, Tech. Rep., June 2021.