MEAN TEACHER MODEL BASED ON CMRANN NETWORK FOR SOUND EVENT DETECTION

Technical Report

Qian Yang *Jing Xia* *Jinjia Wang*

College of Information Science and Engineering,
Yan shan University, Qinhuangdao, China

ABSTRACT

This paper proposes an improved mean teacher model for sound event detection tasks in a domestic environment. The model consists of CNN network, ML-LoBCoD network, RNN network and attention mechanism. To evaluate our method, we tested on the D-CASE 2019 Challenge Task 4 dataset. The results show that the average score of F1 in the evaluation 2018 dataset is 22.7%, and the F1 score in the validation 2019 dataset is 23.4%.

***Index Terms*—** DCASE2019, Sound Event Detection, Mean Teacher, ML-LoBCoD-Net

1. INTRODUCTION

Sound Event Detection (SED) is the automatic recognition of specific sound tasks in a continuous recording. The purpose of sound event detection is to identify sound events in an audio recording, including estimating the start and offset of sound events and giving a label for each event[1].The task of DCASE 2019 task4 is to detect sound events in a home environment. The goal of this task is to evaluate the system used to detect sound events using weakly labeled or unmarked real data and strongly labeled synthetic data (with time stamps) [2].In this paper, the mean teacher model is used to solve the problem of sound event detection. The model consists of CNN network, ML-LoBCoD network, RNN network and attention mechanism. The CNN network and the ML-LoBCoD network are merged networks, which can jointly extract features. The output of the CNN network and the ML-LoBCoD network is averaged and then fed to the RNN network, and then passed to the attention mechanism to obtain the prediction of the final tag class. The mean teacher model is a method of average model weight rather than label prediction, which improves the accuracy of the test, the learning speed and the classification accuracy of the training network[3].

2. METHOD

The CMRANN network model is shown in Figure 1. The input is a log-Mel spectrum and the output is a prediction of the clip category and timestamp. The network structure includes CNN network, ML-LoBCoD-Net with T iteration expansion, RNN network module and attention mechanism module. The CNN network and the ML-LoBCoD network jointly extract features, and the output of the CNN network and the output of the ML-LoBCoD network are averaged and then transmitted to the RNN network. The output of the RNN network is delivered to the attention mechanism network. The attention mechanism can increase the focus on important time frames by weighting, and can automatically select and participate in important frames of the target while ignoring irrelevant parts (such as background noise segments).

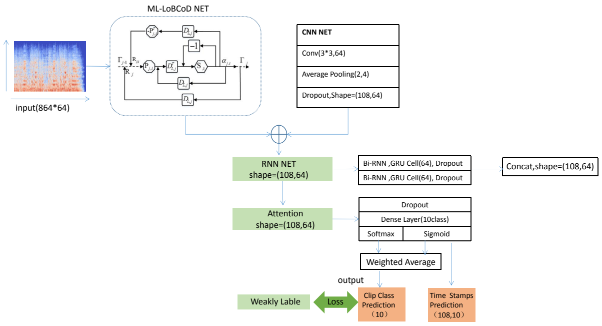

Figure 1: CMRANN Model.

This paper uses a three-layer ML-LoBCoD-Net model with four iterations. ML-LoBCoD-Net is a forward transfer in neural networks. The parameters of the network are exactly the same as the traditional forward transmission type. It can improve the performance of typical CNN without introducing any parameters in the model. ML-LoBCoD-Net firstly uses a simple recursive hierarchical operation to treat the convolutional sparse coding in the slice-based local block coordinate descent method (LoBCoD) as a layer of neural network and expand it into multiple layers.The multi-layer slice-based local fast coordinate descent method (ML-LoBCoD) is then iteratively expanded.

3. MEAN TEACHER

This paper uses the CMRANN network as student model and teacher model. The predictive labels output by the student model and the teacher model are calculated for consistency loss, including strong consistency loss and weak consistency loss, which is mainly to ensure that the prediction result of the teacher model is as similar as possible to the prediction label of the student model. Since the parameters of the teacher model are the average of the student model parameters, the prediction label should not have too much jitter, which is equivalent to smoothing the label to ensure that the output is more stable. The three loss functions are weighted and the student model parameters are updated using a backpropagation algorithm. The teacher model does not directly participate in backpropagation.

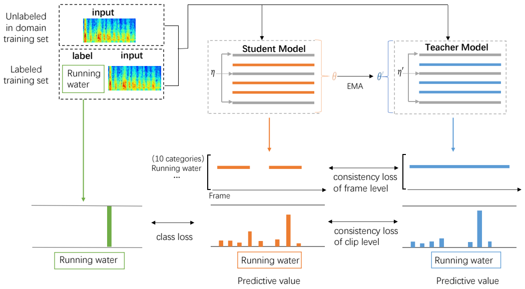

Figure 2: Schematic diagram of the mean teacher model for weakly labeled semi-supervised.

After the student model updates the parameters, the teacher model parameters are updated to the average of the student model parameters.When unlabeled data is trained, only consistency loss is used and no classification loss is used. Both model outputs are available for prediction, but at the end of the training, teacher model predictions are more likely to be correct.

4. EXPERIMENT

4.1. Dataset

The DCASE 2018 task4 data set consists of a 10-second audio clip recorded or synthesized in a domestic environment to simulate a domestic environment. It is a subset of Audiooset provided by Google, consisting of 10 types of sound events, including human and animal sounds, instrument sounds, and common everyday environmental sounds. The training set in the development set includes 1578 weakly labeled audio clips, 1412 unlabeled audio clips in the domain and 2045 strong labeled synthetic audio clips; The validation set consists of 1168 strongly labeled audio clips that are a fusion of the DCASE 2018 Task 4 test set and evaluation set. The evaluation set includes 13190 unlabeled audio clips.

4.2. Setup

Our experiments used logarithmic Mel filtering to process audio clips. Each audio clip is first resampled at 44.1KHZ, and we believe that resampling at low frequencies may confuse categories likeelectric shaver/toothbrush"and "vacuum cleaner".After resampling, a short-time Fourier transform is performed to obtain a spectrogram; then multiplied by a Mel filter group which is a 64-band, and a logarithm is obtained to obtain a logarithmic Mel spectrogram. For each audio clip, you get a 864*64 feature vector.

In the training phase, we use Adam as the optimizer with a learning rate of 0.001. The batch size is 24 and the number of iterations is 100 rounds. The synthetic data sets in the development set are divided into training data and verification data by 80% and 20%. The validation set is used for hyperparameter adjustment and to screen the final model. The validation set with strong tag data in the development set is used as test data for the development set to evaluate the performance of the model in the development set. The evaluation set data is used to make a final assessment of the model.

4.3. Results

Table1 F-score metrics (macro averaged)of evaluation2018 dataset

$class$	$Event-based$	$Segment-based$
$Alarm/bell/ringing$	41.9%	66.0%
$Blender$	27.4%	47.9%
Cat	26.2%	43.8%
$Dishes$	12.4%	31.6%
Dog	5.1%	40.5%
$Electricshaver$	25.0%	59.5%
$Frying$	12.8%	45.7%
$Runningwater$	9.6%	41.1%
$Speech$	36.0%	77.4%
$Vacuumcleaner$	30.2%	60.7%
$Mean$	22.7%	51.5%

Table2 F-score metrics (macro averaged)of validation2019 dataset

$class$	$Event-based$	$Segment-based$
$Alarm/bell/ringing$	38.5%	69.4%
$Blender$	24.8%	47.6%
Cat	29.8%	46.3%
$Dishes$	12.6%	32.6%
Dog	4.7%	37.9%
$Electricshaver$	24.9%	64.4%
$Frying$	13.6%	51.0%
$Runningwater$	12.1%	49.3%
$Speech$	37.2%	78.2%
$Vacuumcleaner$	36.1%	65.8%
$Mean$	23.4%	54.3%

Table 1 and table 2 respectively show the F score of each sound event on the evaluation2018 dataset and validation2019 dataset, as well as the average F1 score of our proposed model.

Table3 Comparition of experimental results

dataset	evaluation 2018		validation 2019	
$result$	$Event$	$Segment$	$Event$	$Segment$
$baseline$	20.6%	51.4%	23.7%	55.2%
$CMRANN-MT$	22.7%	51.5%	23.4%	54.3%

5. REFERENCES

[1] B. McFee, J. Salamon, and J. P. Bello, "A sample paper in conference proceedings," in *Proc. IEEE Transactions on Audio, Speech, and Language Processing*, 2018, pp. 2180–2193.

[2] http://www.Dcase.community/challenge2019/task-sound-event-detection-in-domestic-environments.

[3] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," pp. 1195–1204, 2017.