MEAN TEACHER CONVOLUTION SYSTEM FOR DCASE 2018 TASK 4

Lu JiaKai

PFU SHANGHAI Co., LTD
46 Building 4~5 Floors, 555 GuiPing Road
XuHui District, Shanghai 200233, CHINA
lu_jiakai.pfu@cn.fujitsu.com

ABSTRACT

In this paper, we present our neural network for the DCASE 2018 challenge's Task 4 (Large-scale weakly labeled semi-supervised sound event detection in domestic environments). This task evaluates systems for the large-scale detection of sound events using weakly labeled data, and explore the possibility to exploit a large amount of unbalanced and unlabeled training data together with a small weakly annotated training set to improve system performance to doing audio tagging and sound event detection. We propose a mean-teacher model with context-gating convolutional neural network (CNN) and recurrent neural network (RNN) to maximize the use of unlabeled in-domain dataset.

***Index Terms*—** Mean-teacher, weakly supervised learning, weak labels, context gating, convolutional neural network

1. INTRODUCTION

It has been three years since the sound event detection has been hold by DCASE. Every year, the challenge is closer to the actual situation, both applying and training. The DCASE 2018 challenge's Task 4 aims to exploring the possibility to exploit a large amount of unbalanced and unlabeled training data together with a small weakly annotated training set, which is a typical semi-supervised learning process.

Other than the task in DCASE 2017, there are only a little amount of weakly labeled date has been provided, beside of a large amount of no-labeled data both in in-domain and out-of-domain. In fact, the strongly labeled data is difficult to collect, and may contain the subjective judgment from human that will bias the model not to produce a general result. In comparison, annotating the audio with a label of whole clip is easier. Collecting the unlabeled data both in in-domain and out-of-domain is equally easy. Compared to the ease of collecting data, the training process is more difficult which meaning the semi-supervised method must be used to exploit the unlabeled data-set.

Is this paper, we propose a sound event detector with context-gating convolutional neural network (CNN) [1] [2] and recurrent neural network (RNN) [3] that can recognize sound event from the fully usage of weakly labeled data and the maximize use of in-domain unlabeled data by a semi-supervised model.

2. DATASET

2.1. DCASE 2018 Task 4 Dataset

The dataset of DCASE 2018 challenge's task 4 has 3 parts in training process: weakly label dataset, unlabeled in-domain dataset, and unlabeled out-of-domain dataset. There are 10 class of sound in dataset that appear in different environments.

The weakly label dataset only contains 1578 audio clips, which is nearly 10% of the whole dataset. The unlabeled in-domain dataset contains 14412 audio clips, which is 10 times the weakly label dataset. The unlabeled out-of-domain dataset contains 39999 audio clips, which is the most.

The signal of audio clip is mono-channel and sampled at 44,100 Hz with a maximum duration of 10 seconds. Every audio clip in domain contain more than one sound event that may partly overlap.

2.2. Audio Preprocessing

First, resample the audio clips at 22,050 Hz, because the high frequency part of sound signal is not useful for event detection in daily life. There are some class such as Vacuum cleaner, Electric shaver and Electric toothbrush, which are closer in frequency, so resampling the audio clips at 16,000 Hz may confuse these audios.

Second, extract the log mel-spectrogram from the audio clips by 128-bin, 2048-window and 365-hop (1683-overlap). After that process, a 10-second audio clip should be convert to a 640-frames float data as the audio feature. For the audio clip is not 10-second long, padding or truncating is used.

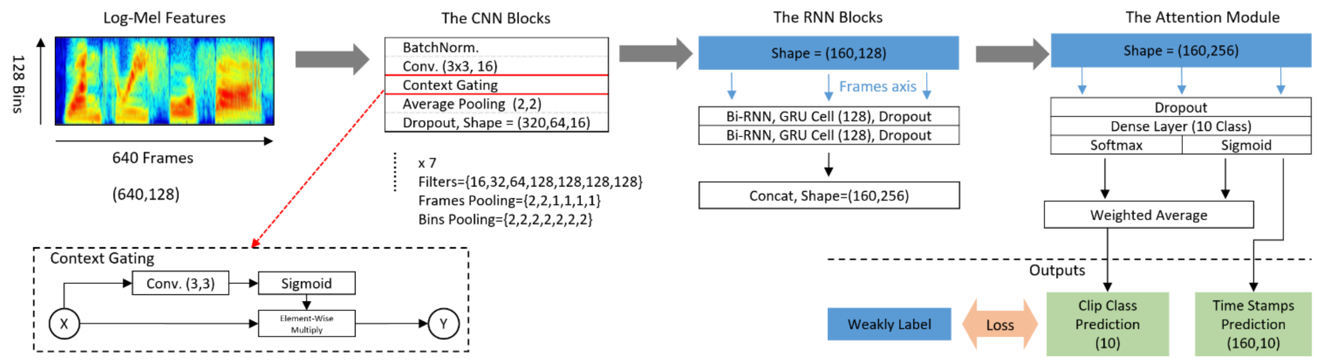

Figure 1: The architecture of the overall neural network. There are 2 final output, one for predicting the location of the sound events and the other one for weakly labeled training.

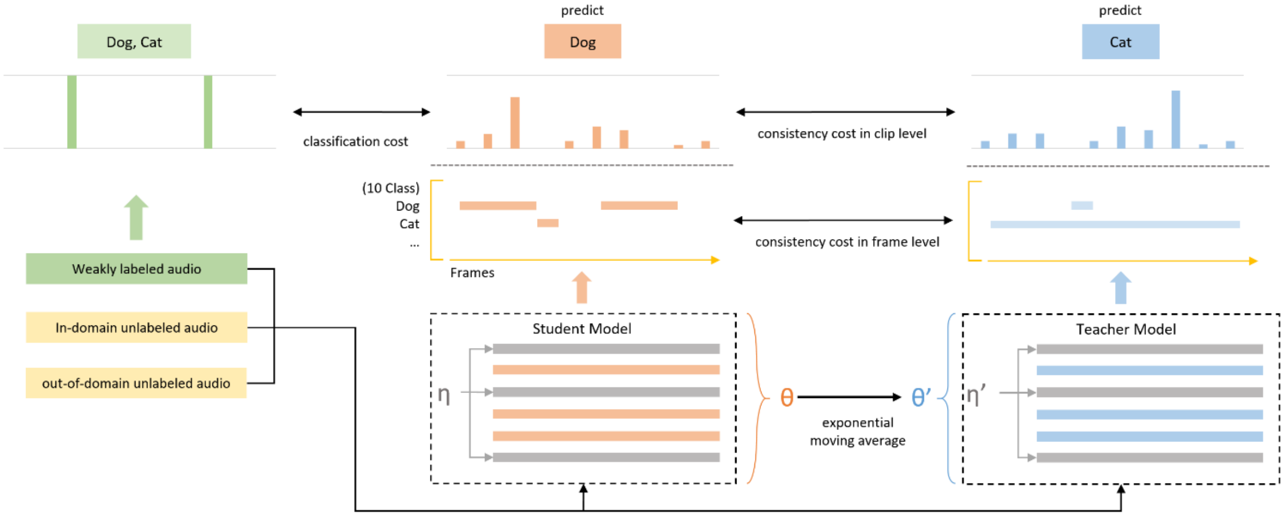
Dog, Cat
predict
Dog
predict
Cat
classification cost
consistency cost in clip level
(10 Class)
Dog
Cat
...
Frames
consistency cost in frame level
Weakly labeled audio
In-domain unlabeled audio
out-of-domain unlabeled audio
Student Model
η
θ
θ'
exponential moving average
η'
Teacher Model

Figure 2: The Mean Teacher method. The figure depicts a training batch with dataset in three types. Both the student and the teacher model evaluate the input applying noise (η; η') within their computation such as dropout. The output of the student model is compared with the multi-label using classification cost and with the teacher output using consistency cost. After the weights of the student model have been updated with gradient descent, the teacher model weights are updated as an exponential moving average of the student weights.

3. PROPOSED METHODS

There are some methods used in model (Figure 1) to improve the performance to detect sound events.

3.1. Context Gating

In order to make the model pay more attention to important parts of audio features in frames axis, we propose to use Context Gating (CG) [4] module as the activation in the CNN part of the model.

$$Y = \sigma(\omega \cdot X + \beta) \odot X \qquad (1)$$

Where $X \in R^n$ is the input feature vector, σ is the elementwise Sigmoid activation and ⊙ is the element-wise multiplication. $\omega \in R^{n \times n}$ And $\beta \in R^n$ are trainable parameters. The vector of weights $\sigma(\omega \cdot X + \beta) \in [0,1]$ represents a set of learned gates applied to the individual dimensions of the input feature X.

The form of the Context Gating layer is inspired by the Gated Linear Unit (GLU) [5], but more efficient by reduces the number of learned parameters.

3.2. Attention Output

Although the Global Average Pooling (GAP) can be presented as an attention model, we use the attention model improved on SURREY-CVSSP SYSTEM [6].

Inspired by the ideas of the Context Gating, the two FNN layers connected with the softmax and sigmoid layer separately will be merged to one FNN layers. The sigmoid as the activation function will do classification at each frame, and the softmax as the activation function will attend the frames that may occur sound event.

The final classification of the audio clip is defined as below:

$$\text{Y}' = \frac{\sum_t^T sigmoid(x) \odot softmax(x)}{\sum_t^T softmax(x)} \qquad (2)$$

Where $X \in R^n$ is the output vector of the merged FNN layers, $\odot$ is the element-wise multiplication. T is final frame-level resolution. There are one tenth scale between the final resolution and the input frames resolution by pooling along the frames axis, it mean that if the input features has 640 frames long, the final T should be 160 frames.

The Y' is the clip-level classification, which can be directly used to make the back-propagate loss by comparing this prediction with the weakly label of the audio clip.

3.3. Mean Teacher

We apply the Mean-Teacher semi-supervised method (Figure 2) [7] to exploit the large amount of unlabeled data effectively. The main purpose of this model is averaging model weights over training steps tends to produce a more accurate model than using the final weights directly.

The teacher model do not participate in the back propagating directly, but use the EMA weights of the student model. There are two loss to calculate out in a training step: classification cost and consistency cost.

The consistency cost in our model is composed of two parts: class consistency in clip-level and in frame-level. Both of them can be obtained by comparing the logits of both the student model and the teacher model for the whole audio clips including labeled and unlabeled.

In the test step, both model outputs can be used for prediction, but at the end of the training the teacher prediction is more likely to be correct.

3.4. Ensemble Model

Since the mean teacher model is the mean of the student model, so the fusion in iterations among one model is not required. We use the mean of the outputs of different models as the fusion model.

4. EVALUATION RESLUT

In DCASE 2018 challenge's task 4, the event-based F1-score is used to evaluate the performances of modules. Additionally, event-based error rate will be provided as a secondary measure.

4.1. Experimental setup

For the single model shown in Figure 1, we use many variation of model to achieve the best performance. There are three parts in the model: CNN Blocks, RNN Blocks and Attention Module. The proposed methods is used in model. The same dropout [8] with 50% rate is used in all layers. We use the Adam-optimizer [9] to accelerate convergence. For the mean teacher, different Consistency cost in the set of 15.0, 10.0, 7.0 and 3.0 is tested, which is the most important parameter of the mean teacher model.

4.2. Results

This section presents the results for the sound event detection on the test set. We use the F1 and ER of micro average as the performance metric.

Models	**F1 (%)**	**ER**
DCASE Baseline	14.06	1.54
Model-noContextGate	24.18	1.21
Model-CG	25.81	1.14
Fusion	**26.36**	**1.09**

Table 1: F1 and ER comparisons of Models without mean teacher method on test set with 33-MedianFilter.

Model (Consistency Cost)	**F1 (%)**	**ER**
DCASE Baseline	14.06	1.54
Mean-Teacher (15.0)	25.92	1.13
Mean-Teacher (10.0)	27.16	1.12
Mean-Teacher (7.0)	27.6	1.08
Mean-Teacher (3.0)	25.58	1.15
Mean-Teacher Fusion	**28.55**	**1.07**

Table 2: Mean teacher models' result on test set with 33-MedainFilter

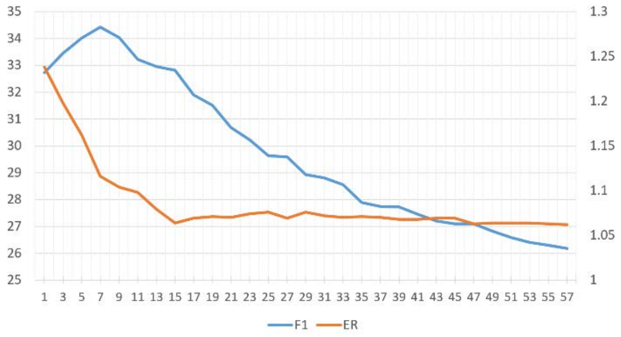

Figure 3: Result in different select of median filter.

Models (Median Filter)	**F1**	**ER**
DCASE Baseline	14.06	1.54
Mean-Teacher Fusion (1)	32.723	1.238
Mean-Teacher Fusion (7)	**34.418**	1.116
Mean-Teacher Fusion (33)	28.55	1.07
Mean-Teacher Fusion (51)	26.587	1.064

Table 3: Result in different select of median filter.

The performances in different select of median filter show is very obvious. By balance of the F1 and ER, we select 7 as Median Filter’s value.

Fusion selections (Consistency Cost)	Median Filter	F1	ER
8.0	33	28.55	1.07
{7.0, 8.0, 9.0, 10.0} Average	33	27.25	1.10
8.0	7	34.42	1.11
{7.0, 8.0, 9.0, 10.0} Average	7	34.30	1.13

Table 4: The 4 predictions with different parameters on Consistency Cost and Median Filter

We submitted four prediction for the same model with different parameters as shown in Table 4. First, in each training step, we select the best 5 iteration on F1-Score to represent the best performance of this training step. Then, we merge these result by averaging the outputs in two types: one is single training step by consistency cost of 8.0, and another is averaging multiple training steps by consistency cost in 7.0, 8.0, 9.0, 10.0. At Final, we use two values of median filter to double the predictions.

5. CONCLUSIONS

In this paper, the mean teacher model with context-gating CNN and Bi-RNN was proposed to exploit a large amount of unbalanced and unlabeled training data together. An error rate of 1.16 and F-score of 34.418% was achieved on the test data. Due to lack of time, there are still potential improvements can be achieve in this models in the future.

6. REFERENCES

[1] K. J. Piczak, “Environmental sound classification with convolutional neural networks,” in Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on. IEEE, 2015.

[2] H. Zhang, I. McLoughlin, and Y. Song, “Robust sound event recognition using convolutional neural networks,” in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE.

[3] G. Parascandolo, H. Huttunen, and T. Virtanen, “Recurrent neural networks for polyphonic sound event detection in real life recordings,” in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016.

[4] A. Miech, I. Laptev and J. Sivic, “Learnable pooling with Context Gating for video classification” in arXiv: 1706.06905, 2017.

[5] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” in arXiv: 1612.08083, 2016.

[6] Y. Xu, Q. Kong, W. Wang, Mark D. Plumbley, “Attention and Localization based on a Deep Convolutional Recurrent Model for Weakly Supervised Audio Tagging” in arXiv: 1703.06052, 2017.

[7] A. Tarvainen, H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results” in arXiv: 1703.01780, 2017.

[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” in Journal of Machine Learning Research (JMLR), 2014.

[9] D. Kingma and J. Ba, “Adam: A method for stochastic optimization”, arXiv: 1412.6980, 2014.