MEAN TEACHER CONVOLUTION SYSTEM FOR DCASE 2020 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 2020 challenge's Task 4 (Sound event detection and separation 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 convolutional neural network (CNN) and recurrent neural network (RNN) to maximize the use of unlabeled in-domain dataset. The architecture is based on our 2018 competition model.

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

1. INTRODUCTION

Compared to DCASE 2018, DCASE2020 adds more secondary goals and data. Due to the limited time period, I only made the main goal, and did not do the sound separation and PSDS statistics. This work is based on my results in the DCASE 2018[1], which improved the model and fixed some defects.

Other than the task in DCASE 2018, the dataset in DCASE 2020 provides synthetic data with strong labels. With the support of strong label data, the model is better trained than ever.

Is this paper, we propose a sound event detector with 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.

Since this is a competition, not a paper, some opinions and methods have strong individual subjectivity. And the correctness of these opinions cannot be guaranteed.

2. DATASET

2.1. DCASE 2020 Task 4 Dataset

The dataset of DCASE 2020 challenge's task 4 has 3 parts in training process: weakly label dataset, synthetic dataset, and unlabeled 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 dataset contains 14412 audio clips, which is 10 times the weakly label dataset. The synthetic dataset contains 2049 audio clips, which is the most. We found that the number of foreground sounds of synthetic data is little, so only the official generated synthetic data is used.

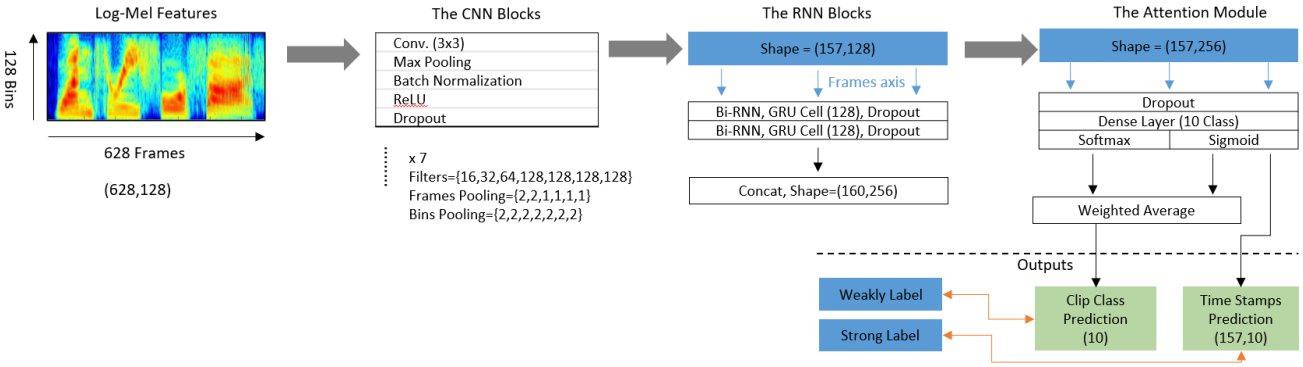
The signal of audio clip is mono-channel and sampled at 44,100 Hz with a maximum duration of 10 seconds. Every audio clip in domain contain more than one sound event that may partly overlap.

2.2. Audio Preprocessing

First, resample the audio clips at 16,000 Hz, because the high frequency part of sound signal is not useful for event detection in daily life. Some experiments prove that 16000 contains enough acoustic features which can make the model converge faster.

Second, extract the log mel-spectrogram from the audio clips by 128-bin, 2048-window and 255-hop. After that process, a 10-second audio clip should be converted to a 628-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.



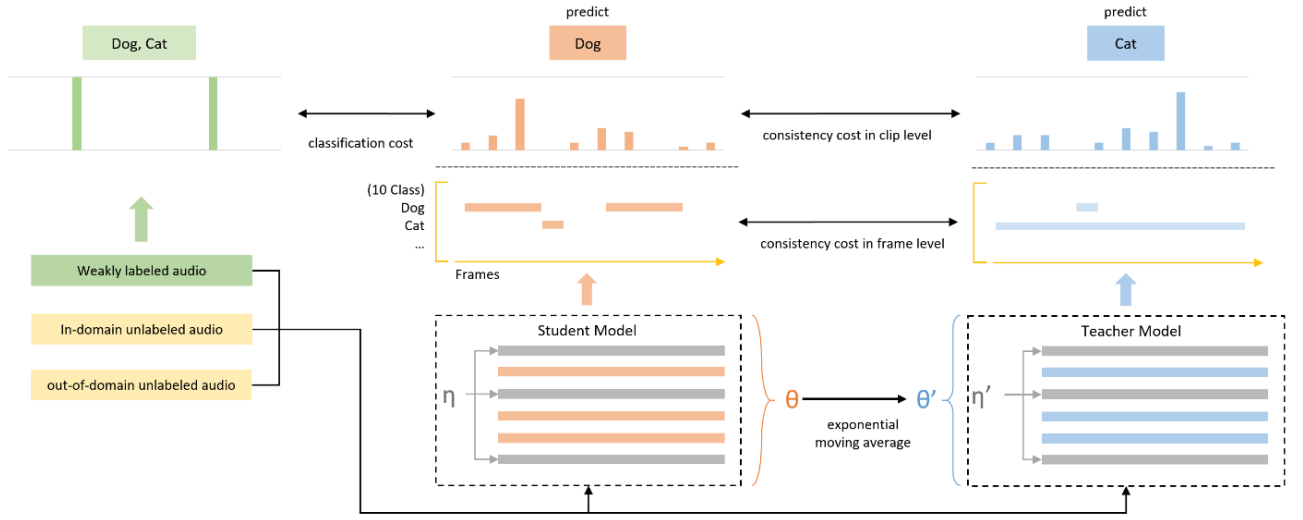


Figure 2: The Mean Teacher method. The figure depicts a training batch with dataset in three types. Both the student and the teacher model evaluate the input applying noise (η; η') within their computation such as dropout. The output of the student model is compared with the multi-label using classification 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. CNN Model

The Context Gating used in 2018 was abandoned. Instead, the standard CNN model is used. According to the benchmark on Batch Normalize Layer of caffenet [10] and the researcher on Batch Normalize [12], we tested some models with different order of layers. Although many papers show that BN works well after the activating function, in our experiment, the order of the following layers achieves the best effect:

Conv→MaxPool→BatchNormalization→ReLU→Dropout

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:

$$Y' = \frac{\sum_t^T sigmoid(x) \odot softmax(x)}{\sum_t^T softmax(x)} \quad (1)$$

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. Circle Loss

We apply a new loss function called Circle Loss[12], It can directly replace the Binary Crossentropy we used before and can slightly improve the accuracy with an ideal extensiveness.

3.4. 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.

3.5. New Consistency Cost

The Guided Learning used in last year's champion model inspired us. We used it to replace the original loss of consistency cost where i is the epoch number during training and γ is a hyper-parameter.

$$\alpha = \begin{cases} 0, & i < start\_epoch \\ 1 - \gamma^{i - start\_epoch}, & otherwise \end{cases} \quad (2)$$

We use $\gamma = 0.999$ in our model.

3.6. 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. At the same time, a category-based fusion method is also used.

4. VALIDATION RESLUT

In DCASE 2020 challenge's task 4, the event-based F1-score is used to evaluate the performances of modules. Due to lack of time, we only used PSDS Score on the evaluation set, not on the validation set.

4.1. Experimental setup

For the single model shown in Figure 1, we use many variations 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 10% rate is used in all layers. We use the Adam-optimizer [9] to accelerate convergence.

4.2. Results

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

Models	F1 (%)
DCASE Baseline	34.80
Model-Single-Best	41.96
Iterations Fusion Top 3	44.06
Iterations Fusion Top 5	44.73
Iterations Fusion Top 7	44.84
Iterations Fusion Top 10	44.59
Class-wise Fusion Top 1	45.91
Class-wise Fusion Top 3	46.41
Class-wise Fusion Top 5	46.58
Class-wise Fusion Top 7	47.51
Class-wise Fusion Top 10	**47.85**

Table 1: F1 comparisons of Models on validation set.

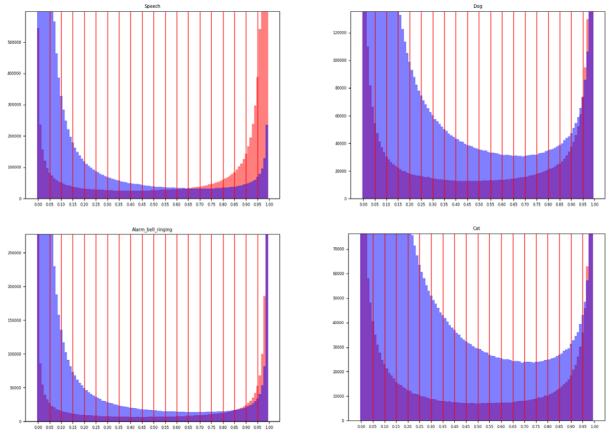


Figure 3: Some classified prediction distributions, red indicates correct and blue indicates error.

5. CONCLUSIONS

In this paper, the mean teacher model with CNN and Bi-RNN was proposed to exploit a large amount of unbalanced and unlabeled training data together. An error rate of 0.90 and F-score of 46.09% was achieved on the test data. Due to lack of time, there are still potential improvements can be achieved in this model 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.

[10] BatchNormalize benchmark in https://github.com/ducha-aiki/caffenet-benchmark/blob/master/batchnorm.md, 2016

[11] X. Li, S. Chen, X. Hu, J. Yang, "Understanding the Disharmony between Dropout and Batch Normalization by Variance Shift" in arXiv: https://arxiv.org/pdf/1801.05134.pdf, 2018

[12] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, Y. Wei, "Circle Loss: A Unified Perspective of Pair Similarity Optimization" in arXiv: https://arxiv.org/abs/2002.10857