AUDIO TAGGING WITH MINIMAL SUPERVISION BASED ON MEAN TEACHER FOR DCASE 2019 CHALLENGE TASK 2

Technical Report

Jun He, Penghao Rao, Bo Sun, Lejun Yu,*

Beijing Normal University, College of Information Science and Technology
No. 19, XinJieKouWai St., HaiDian District, Beijing 100875,China
hejun@bnu.edu.cn, 201721210016@mail.bnu.edu.cn, {tosunbo, yulejun}@bnu.edu.cn

ABSTRACT

In this report, we describe the mean teacher based audio tagging system and performance applied to the task 2 of DCASE 2018 challenge, where the task evaluates systems for audio tagging with noisy labels and minimal supervision. The proposed system is based on a VGG16 network with attention mechanism and gated CNN. Following data augmentation techniques are used to increase model robustness: a) Scaling the signal with 0.75 to 1.5 time, b) Adding Gaussian white noise with 20dB to 40dB. Samples with noisy labels are regarded as unlabeled and are utilized with semi-supervision method namely mean teacher. The proposed system is trained using 5-fold cross-validation, and the final result is the arithmetic mean of the five models. Finally, the method provides lwlrap score of 0.631, which is measured through the Kaggle platform.

Index Terms— Audio-tagging, mean teacher, convolutional neural networks, attention mechanism,

1. INTRODUCTION

In recent years, computer vision techniques such as image classification, object detection, etc. have gained lots of significant achievements benefits from deep learning and comprehensive datasets. Similarly, audio tagging task may benefit a lot from large scale sound datasets. Google researchers developed AudioSet[1], a dataset and ontology of audio events that endeavors to provide comprehensive coverage of real-world sounds at ImageNet-like scale. AudioSet is the largest audio dataset available so far. However, even though the dataset annotations have been manually validated, about 15% of the categories present a quality estimate with a score below 50%. Freesound dataset (FSD)[2] is also a large-scale, general-purpose audio dataset under development that is composed of Freesound content annotated with labels from the AudioSet Ontology. In DCASE 2018 challenge[3], participants need to predict one out of 41 classes for general-purpose audio tagging using a reduced subset of FSD. This year, the challenge[4] is took to the next level with multi-label audio tagging, doubled number of audio categories, and a noisier than ever training set containing about 5000 manually-labeled samples, and about 20000 samples with a certain amount of noise.

In our submission, we regardless the labels of noisy subset. In this way, we turn the problem into semi-supervised problem which is similar to DCASE 2018 Task 4 but with more categories and no need to locate audio events. Mean teacher is proved a good method to deal with semi-supervised problems and is the backbone of our system. Besides, we utilize data augment, attention mechanism and gated CNN to improve the performance. The detailed of our submission is presented following.

*Corresponding Author

2. PROPOSED METHOD

2.1. Data augmentation and feature extraction

We remove the potential silence in the beginning and the end of the audio signal. To increase system robustness, we augment data with gaussian white noise with SNR from 20 to 40dB and scale the signal length from 0.75 to 1.5 times. Then we randomly pick 1 second out of the audio clip and extract the mel-spectrogram by 128 bins, window size of 0.025 second and hop size of 0.010 second. Zero padding is applied for some clips that are less then 1 second. We also calculate the delta and accelerate of mel-spectrogram since they may capture the dynamic information in signal and concatenate the features at channel dimension. Finally, we obtain a 3D feature representation X^{t*f*c}, where t is 100 in our system, representing the number of frames, f is 128, the bins of mel-spectrogram, c is 3, representing mel-spectrogram, delta and accelerate of mel-spectrogram respectively.

2.2. Proposed network

The prototype of our network is VGG16, with gated CNN and attention mechanism added and some hyperparameters changed[5]. The model comprises about 20.7M trainable parameters.

2.2.1. Gated CNN

We apply Context Gating modules[6] at begin of every CNN blocks in the model to capture the important parts in feature maps.

$$Y = \sigma(\omega \cdot X + \beta) \odot X \tag{1}$$

where X is the input feature vector, σ is the element-wise sigmoid activation, ω and β are trainable parameters, $\odot$ is element-wise multiplication. The vector of σ represent the importance of input X. The mechanism can be implemented simply by the element-wise multiplication of sigmoid activated convolution layer and another normal convolution layer.

Input batchsize*100*128*3
3*3 Gated Conv(pad-1, stride-1)-64-BN-LeakyReLU
3*3 Conv(pad-1, stride-1)-64-BN-LeakyReLU
2*2 Max-Pooling
3*3 Gated Conv(pad-1, stride-1)-128-BN-LeakyReLU
3*3 Conv(pad-1, stride-1)-128-BN-LeakyReLU
2*2 Max-Pooling
3*3 Gated Conv(pad-1, stride-1)-256-BN-LeakyReLU
3*3 Conv(pad-1, stride-1)-256-BN-LeakyReLU
3*3 Conv(pad-1, stride-1)-256-BN-LeakyReLU
2*2 Max-Pooling
3*3 Gated Conv(pad-1, stride-1)-512-BN-LeakyReLU
3*3 Conv(pad-1, stride-1)-512-BN-LeakyReLU
3*3 Conv(pad-1, stride-1)-512-BN-LeakyReLU
2*2 Max-Pooling
3*3 Gated Conv(pad-1, stride-1)-512-BN-LeakyReLU
3*3 Conv(pad-1, stride-1)-512-BN-LeakyReLU
3*3 Conv(pad-1, stride-1)-512-BN-LeakyReLU
2*2 Max-Pooling
Attention
Dense(1024)-BN-LeakyReLU
Dense(512)-BN-LeakyReLU
80-way Softmax

Table 1: Network architecture

2.2.2. Attention mechanism

Since there are some silent frames and some features carrying no useful information, it is necessary score the importance of features. Assuming the output of last CNN block is X^{t*f*c} and regard X^{t*f*c} is composed of $x_\tau^{\hat{f}}$ with t steps, where $\hat{f} = f * c$. The proposed attention mechanism can be formulated as following:

$$\begin{aligned} \alpha_\tau &= \frac{e^{u^T x_\tau^{\hat{f}}}}{\sum_\tau^t e^{u^T x_\tau^{\hat{f}}}} \\ z &= \sum_{\tau=1}^{t} \alpha_\tau y_\tau \end{aligned} \tag{2}$$

where u is attention weight, α_τ is the important score of feature $x_\tau^{\hat{f}}$.

2.3. Mean teacher

Since we regardless the labels of noisy subset, the task becomes a semi-supervised learning problem where mean teacher method is very suitable. Mean teacher[7] build a teacher model and a student model where teacher model average the weights of student model to generate predictions and don't propagate gradient directly during train steps. The output of student model and teacher model can be used for prediction, but the prediction of teacher model is more likely to be correct. The cost of mean teacher is composed of classification loss and consistency loss. The coefficient of exponential moving average and consistency loss increase with the increase of training epoch.

3. SUBMISSION AND RESULTS

Our submission get LB lwlrap 0.631 on Kaggle. We also tried a supervised system which get LB lwlrap 0.611, with curated data only and no data augmentation. Semi-supervised method is slightly better than supervised one. However with more fine-tuned hyperparameters, semi-supervised system may get better result.

4. CONCLUSION

The report described a system submitted to DCASE 2019 Task2, in which the task is regarded as semi-supervised problem. A few steps are involved in mel-spectrogram feature extraction including slience removal and data augment. The proposed model is based on a VGG16 network with attention mechanism and gated CNN and trained with mean teacher method. The lwlrap of the system is 0.631 and may get better with fine-tuned hyperparameters. Other semi-supervised methods such as MixMatch are also worth trying.

5. REFERENCES

[1] Gemmeke, Jort F., Ellis, Daniel P. W., Freedman, Dylan, Jansen, Aren, Lawrence, Wade, Moore, R. Channing, Plakal, Manoj and Ritter, Marvin "Audio Set: An ontology and human-labeled dataset for audio events." in *ICASSP , IEEE* , pp. 776-780 , 2017.

[2] Fonseca, E.; Pons, J.; Favory, X.; Font, F.; Bogdanov, D.; Ferraro, A.; Oramas, S.; Porter, A. and Serra, X. " Freesound Datasets: A Platform for the Creation of Open Audio Datasets.", in *ISMIR* , pp. 486-493 , 2017.

[3] Kong, Qiuqiang and Iqbal, Turab and Xu, Yong and Wang, Wenwu and Plumbley, Mark D., "DCASE 2018 Challenge baseline with convolutional neural networks.", in *CoRR* abs/1808.00773 (2018)

[4] Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, and Xavier Serra. "Audio tagging with noisy labels and minimal supervision". Submitted to *DCASE2019 Workshop*, 2019. URL: https://arxiv.org/abs/1906.02975

[5] L. JiaKai, "Mean teacher convolution system for dcase 2018 task 4" *DCASE2018 Challenge*, Tech. Rep., September 2018.

[6] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *arXiv* 1612.08083, 2016.

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