MULTI-LABEL AUDIO TAGGING WITH NOISY LABELS AND VARIABLE LENGTH

Technical Report

Boqing Zhu[1], *Kele Xu*[1], *Dezhi Wang*[2], *Mathurin ACHE*[3]

[1] National University of Defense Technology, College of Computer, Changsha, China,
[2] National University of Defense Technology, College of Meteorology and Oceanography, Changsha, China,
[3] Orange Telecom, Paris, France
zhuboqing09@nudt.edu.cn

ABSTRACT

This paper describes our approach for DCASE 2019 Task2: *Audio tagging with noisy labels and minimal supervision*. This challenge uses a smaller set of manually labeled data and a larger set of noise-labeled data to enable the system to perform multi-label audio tagging tasks with minimal supervision conditions. We aim to tagging the audio clips with convolutional neural network under a limited computation and storage resources. To tackle the problem of noisy label data, we propose a data generation method named Dominate Mixup. It can restrain the impact of incorrect label during back propagation and it's suitable for multi-class classification problem. In response to the variable length of audio clips, we conduct an efficient learning method with cyclical audio length which allow us to learn more pattern from widely diverse sound events. On the public leaderboard for the competition, our single model and simple ensemble of 5 models score 0.711 and 0.725 respectively.

Index Terms— Audio tagging, DCASE 2019, CNN, Noisy data, Variable length

1. INTRODUCTION

This paper describe our approach to the Freesound Audio Tagging 2019 which is carried out as Task2 of DCASE 2019 Challenge: Audio tagging with noisy labels and minimal supervision [1]. We have to experience the vastness of sounds on web, mobile devices and sensors, the demand for automatic general-purpose audio tagging systems is increasing dramatically. Recently, deep learning has become the mainstream method of building the systems. In this paper, we present an efficient multi-label audio tagging system based on convolutional neural network.

The primary motivation for this challenge is to predict foster research towards more general machine listening systems capable of recognizing and discerning a wide range of acoustic events and audio scenes. This challenge uses a smaller set of manually labeled data [2] and a larger set of noise-labeled data [3] to enable the system to perform multi-label audio tagging tasks. This task will provide insight towards the development of broadly-applicable sound event classifiers able to cope with label noise and minimal supervision conditions.

Under a limited computation and storage resource, we mainly focus on adequately exploiting the noisy labeled data and dealing with the variable audio length. Firstly, to tackle the problem of noisy labeled data, we propose a novel data generation method named Dominate Mixup. It can restrain the impact of incorrect label during back propagation and it suits well with the multi-class classification problem. Secondly, although the length of the audio is variable, most of the previous work [4, 5, 6] have chosen a empirical length of the audio (for example, 1.5s) to feed in the network. However, we found that the sound events in the audio clip can be divided into two types. One is transitory and repeated, another is persistent and long-lasting. For different sound event, the network need different length of audio to recognize them. Therefore, we conduct a Cyclical Audio Length Schedule to deal with this problem. Our best score on the public leaderboard is 0.725 with an ensemble of 5 models. It is worth noting that with our innovative approach, our single model can achieve 0.711 lwlrap score.

2. DATA PREPROCESSING AND AUGMENTATION

Log-scaled Mel-spectrograms (Log-mel) have shown good performance in the audio related learning system [7, 3]. Thus, we use Log-mel as input of our approach. Firstly, silence parts of at the start and end positions of the audio have been removed. But the silence part between the sound event have been retained, because we hope to keep the original sequential information of the audio. Then, we transform the audio to log-mel using LibROSA [8]. After several comparative tries, 40ms frame width and 5ms frame shift have been selected in the FFT transformation. We keep 128 mel bands in the mel-spectrograms. We directly use 44.1KHz as the sample rating because it reserve more high frequency information and resample the audio will take much time which is expensive in the time limit of the challenge. The preprocessing of curated data and noisy data is the same.

Insufficient labeled data often make the model tend to over fitting, and data augmentation is common used for deep network training. We use SpecAugment [9] to enrich our training data. The augmentation policy include frequency masking and time masking. Frequency masking is applied so that f consecutive mel frequency channels $[f_0, f_0 + f)$ are masked, where f is first chosen from a uniform distribution from 0 to the frequency mask parameter F, and f_0 is chosen from $[0, \nu - f)$. ν is the height of the input Log-mel feature. We set $F = 10\% \times \nu$, which at most 10% mel-bands will be masked. Time masking is applied so that t consecutive time steps $[t_0, t_0 + t)$ are masked, where t is first chosen from a uniform distribution from 0 to the time mask parameter T, and t_0 is chosen from $[0, \tau - t)$. τ is the width of the input Log-mel feature. We set $T = 20\% \times \tau$, which at most 20% frames will be masked.

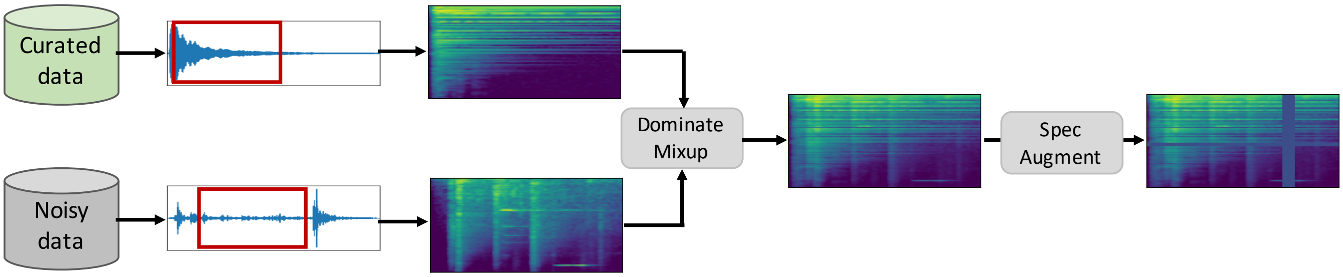

Figure 1: Training data generation pipeline.

3. METHODS

3.1. Network Architecture

Our neural network is based on MobileNet v2 [10] which significantly decreasing the number of operations and memory needed while retaining the same performance. It mainly benefits from the inverted residual with linear bottleneck. The convolutional module takes as an input a low-dimensional compressed representation which is first expanded to high dimension and filtered with a light weight depth-wise convolution. Features are subsequently projected back to a low-dimensional representation with a linear convolution. In the classifier module, we set up 80 units to match the number of audio classes. The network contains totally 18 convolutional layers. After all the convolutional layers, we set up a Adaptive Max Pooling layer which can handle the input feature maps with different sizes. Our modified MobileNet v2 contains a total of 2.3M parameters.

3.2. Dominate Mixup

The given dataset in the task contains two parts, a small set of manually-labeled data which we could consider the label to be curated, and a larger set of noisy-labeled data. The label of noisy-labeled data is making predictions with pre-trained models, these automatically inferred labels might include a substantial level of Incorrect labels. A key focus of this task is to find an appropriate way to adequately exploit a small amount of reliable, curated data, and a larger quantity of noisy data.

We propose a method named Dominated Mixup to make full use of curated and noise data while suppressing the effects of incorrect labels. Unlike the past work using MixUp [11, 12, 13], we guarantee that at least one curated data is involved at every mix process. At the same time, ensure that the curated data accounts for a larger proportion of the mixed data. For a pair of example with their corresponding one-hot label (x_i, y_i), (x_j, y_j), the Dominate Mixup computing (x'_i, y'_i) by

$$\lambda \sim Beta(\alpha, \alpha) \tag{1}$$

$$\lambda' = max(\lambda, 1 - \lambda) \tag{2}$$

$$x_i \in \mathcal{X}_c, x_j \in \mathcal{X}_n + \mathcal{X}_n \tag{3}$$

$$x'_i = \lambda' x_i + (1 - \lambda')x_j \tag{4}$$

$$y'_i = \lambda' y_i + (1 - \lambda')y_j \tag{5}$$

Where λ obeys the Beta distribution with α as a parameter, $\mathcal{X}_c$ is the curated data set while $\mathcal{X}_n$ is the noisy data set. For each x_i in $\mathcal{X}_c$, we randomly pich an x_j from $\mathcal{X}_c + \mathcal{X}_n$. In the experimental part, we compare the results of x_j are picked only from $\mathcal{X}_n$ or $\mathcal{X}_c$.

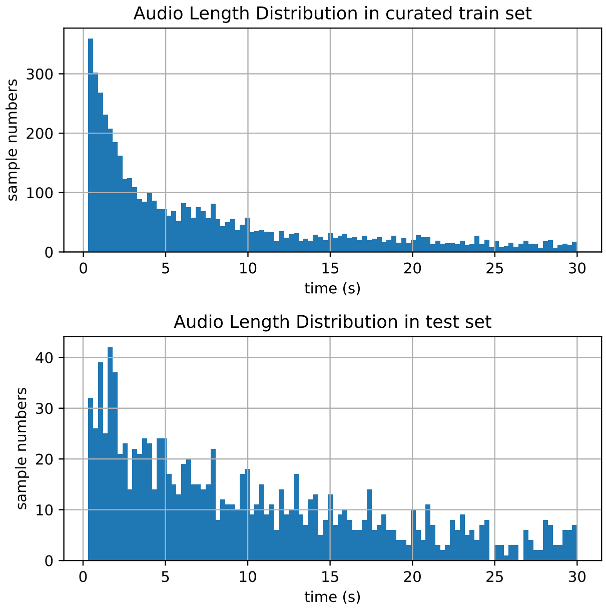

Figure 2: Audio Length Distribustion.

Eq. 2 makes x_j a small proportion of the mixed samples, which can weaken the influence of the incorrect label in backpropagation. It should be noticed, the task is a multi-class classification. As mixup method adds one-hot labels, it is well suited for multi-class classification.

During the test stage, we make predictions without any mixup and any data augmentation. We crop a dozen of Log-mel feature with multiple audio length (3s, 4s, 5s). Input the cropped feature into the trained model and (arithmetic) average the output logits.

3.3. Cyclical Audio Length

In the general audio tagging, the length vary of audio clips is a common and intractable problem we have to face. Fig.2 shows the distribution of the length of the audio clip. After attentive observation of the audio data, we found that the sound events in the

Table 1: Important hyperparameters and values

Hyperparameter	Value
α in Dominate Mixup	1
Batch size	32
Initial learning rate	1e-04
Eta min	1e-06
Epoch	200
Weight decay	5e-06

audio clip can be divided into two types. One is transitory and repeated, like *Bark, Church_bell, Clapping, Knock and Meow*. Another kind of sound event is persistent, long-lasting, like *Accelerating_and_revving_and_vroom, Acoustic_guitar, Car_passing_by, Fill_with_liquid and Raindrop*. Some sound event may take less than one second to recognize the classes of the sound, but others may takes 4-5 seconds to complete. Therefore, we employ a Cyclical Audio Length Schedule to deal with this problem.

We periodically change the audio length feed in the network of each epoch in the network training. The audio length D can be got by

$$D = \frac{1}{2}(D_{max} - D_{min})\sin(2\pi\frac{T_{cur}}{T_{max}}) + \frac{1}{2}(D_{max} + D_{min}) \quad (6)$$

Where D_{max} and D_{min} is the max and min audio length feed in the network, T_{max} is the epoch period of variation, T_{cur} is the current epoch number. In our experiment, we set $D_{max} = 5$, $D_{min} = 2$ and $T_{cur} = 10$. If the length of an audio clip is longer than the window length D, we randomly crop the Log-mel feature of length D in the time dimension. If the length of an audio segment is shorter than the D, we randomly fill 0 on both sides of the Log-mel feature.

Cyclical Audio Length Schedule has two advantages. First, it deals with the problem of different audio lengths, so that the model learns more pattern from different lengths clips. At the same time, in the case of the same average length, there is no increase in the amount of computation. Secondly, it played the role of regularization which makes it easier for the weights in the deep neural network to get rid of minimum values during training. In addition to audio tagging, this schedule can be applied to similar problems with other variable length input.

3.4. Details

We create a criterion between the target y and the output $\hat{y}$ by Binary Cross Entropy:

$$\ell(x, y) = \{l_1, l_2, ..., l_N\} \quad (7)$$

$$l_n = -[y_n \cdot \log\sigma(\hat{y_n}) + (1 - y_n) \cdot \log(1 - \sigma(\hat{y_n}))] \quad (8)$$

Where N is the batch size, σ is a Sigmoid operation. We use 5-folds cross validation to make the results more stable in our experiments. All the models were trained from scratch. After selecting an initial learning rate, we use the warmup method to linearly increase the initial learning rate by 10 times within 5 epochs, and use cosine annealing schedular to gradually reduce the learning rate to Eta_min during the subsequent training. Important hyperparameters are given in the Table.1. We have trained the model for about 1.5 hours (7.5 hours for 5-folds cross models) on a GeForce RTX 2080Ti GPU.

Table 2: Effect on Cyclical Audio Length Schedule

Audio Length	lwlrap in Public Leaderboard
$D = 2$	0.662
$D = 3$	0.676
$D = 4$	0.670
Cyclical Audio Length Schedule	0.689

Table 3: Effect on Dominate Mixup

Method	lwlrap in Public Leaderboard
Dominate Mixup	0.711
Dominate Mixup with $x_j \in \mathcal{X}_n$	0.709
Dominate Mixup with $x_j \in \mathcal{X}_c$	0.705

4. RESULTS

For this challenge, we submitted 3 prediction results. The two of them (Boqing_NUDT_task2_1 and Boqing_NUDT_task2_2) are ensembles of 5 different models with variable hyperparameters. These two submission got 0.725 and 0.723 lwlrap scores in the public leaderboard.

The last submission (Boqing_NUDT_task2_3) is a single model submission which pay more attention to innovative methods without losing much score. It got 0.711 lwlrap scores in the public leaderboard. Table.2 and Table.3 demonstrate the effectiveness of Cyclical Audio Length Schedule and Dominate Mixup respectively. Note that, the results in the Table.3 use the Cyclical Audio Length Schedule we proposed.

5. CONCLUSION

Collecting reliable labeled data is expensive and time-consuming for supervised learning tasks, while data automatically labeled by pre-trained models might include a substantial level of Incorrect labels. The experimental results presented in this paper show that it is possible to increase the performance by employ the noisy labeled data by the Dominate Mixup. This method can achieve a better performance compared with using curated data only. Also, it makes the network more robust. Audio clips length vary is an audio specific problem. For this problem, we proposed the Cyclical Audio Length Schedule to make the network learns more pattern from different length audio clips. Experimental results show that it can achieve better performance when the average audio length is the same. On the submitted results, our approach shows a significantly improvement compare to the baseline system.

6. REFERENCES

[1] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," *arXiv preprint arXiv:1906.02975*, 2019.

[2] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, Suzhou, China, 2017, pp. 486–493.

[3] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," in *Proc. IEEE ICASSP 2019*, Brighton, UK, 2019.

[4] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Procedia Computer Science*, vol. 112, pp. 2048–2056, 2017.

[5] Y. Xu, Q. Huang, W. Wang, and M. D. Plumbley, "Hierarchical learning for dnn-based acoustic scene classification," *arXiv preprint arXiv:1607.03682*, 2016.

[6] B. Zhu, C. Wang, F. Liu, J. Lei, Z. Lu, and Y. Peng, "Learning environmental sounds with multi-scale convolutional neural network," *arXiv preprint arXiv:1803.10219*, 2018.

[7] 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, pp. 1–6.

[8] LibROSA. [Online]. Available: https://librosa.github.io/librosa/

[9] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[11] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[12] V. Verma, A. Lamb, C. Beckham, A. Courville, I. Mitliagkis, and Y. Bengio, "Manifold mixup: Encouraging meaningful on-manifold interpolation as a regularizer," *stat*, vol. 1050, p. 13, 2018.

[13] K. Xu, D. Feng, H. Mi, B. Zhu, D. Wang, L. Zhang, H. Cai, and S. Liu, "Mixup-based acoustic scene classification using multi-channel convolutional neural network," *arXiv preprint arXiv:1805.07319*, 2018.