STACKED CONVOLUTIONAL NEURAL NETWORKS FOR AUDIO TAGGING WITH NOISE LABELS

Technical Report

Yanfang LIU , Qingkai WEI

Beijing Kuaiyu Electronics Co., Ltd., Beijing, PRC.
{liuyf, wqk}@kuaiyu.com

ABSTRACT

This technical report describes the system we used to participate in task 2 of the DCASE 2019 challenge. The task is to predict the tags of audio recordings with using a small number of manually-verified labels and a much larger number of noisy labels. In this task, we propose serveral convolutional neural networks to learn from log-mel spectrogram features. To improve the performance, different techniques preprocessing, data augmentations, loss functions and cross-validation are involved. The prediction results are then ensembled using geometric mean. On the test set used for evaluation, our system achieved a score of 0.734.

Index Terms— predict the tags of audio, noisy labels, convolutional neural networks, data augmentation, ensembled

1. INTRODUCTION

Audio classification is to identify the class of sounds in a given audio signal, where the classes to be detected are typically defined in advance. The Detection and Classification of Acoustic Scenes and Events (DCASE) [1] is a recurring challenge with several tasks pertaining to the classification of audio. This paper describes the system we used to participate in DCASE 2019, task 2 [2].

In task 2, participants are provided with a dataset developed by the Freesound initiative [3]. The training part consists of manually-labeled (curated) data from Freesound Dataset (FSD) and noisy-labeled data from Yahoo Flickr Creative Commons 100M dataset (YFCC). The testing data consists of audio from FSD dataset. The sampling rate of audio recordings are all 44.1 kHz. There are 80 sound event classes in the vocabulary and each audio clip is associated with more than one class.

In recent times, the state-of-the-art in machine learning has come from research in neural networks [4]. It's of the same situation in audio classification, with many of the top submissions in DCASE challenges utilizing neural network architectures [5, 6, 7]. We follow this trend and use two network structure: one is InceptionV3 for training curated data, the other is simple cnn model with 8 layers for training all data. The purpose of training multiple models is to use an ensembling method to combine their abilities of prediction. To achieve this, we use a popular technique called geometric mean.

The key point of task 2 is the subset with noisy labels. The noisy labels can negatively affect the performance of the system and should be handled appropriately. We have tried new Google semi-supervised method, Mixmatch[8], however it didn't work well for the task till now. Final we use some label smoothing methods to improve our model trained with noisy labels.

Other than the noisy labels, there are lots of other tips useful to be demonstrated. Firstly, the lengths of the audio clips vary greatly, from 0.3 s to 30 s. This is a problem because basic convoluntional neural network models expect a fixed-size feature as input. Another consideration is the silence of the inputs; does the entire clip contain important information or only specific part? We address both these issues later in the paper and present our findings.

The rest of this paper is organized as follows. In Section 2, the preprocessing and feature extraction methods are described. In Section 3, the neural network architectures, training methodology, and ensembling algorithm are presented. The results are then given in Section 4. Finally, we summarize in Section 5.

2. PREPROCESSING AND FEATURE EXTRACTION

Prior to training, we applied preprocessing to the inputs followed by feature extraction. Our preprocessing step consists of silent removal (60 dBFS) and zero-padding to the beginning and end of short data. Then we use mel-spectrogram as input feature. Mel-spectrogram was used by most of the top teams in DCASE 2018 challenge and considered to be most suitable feature for audio tagging task. The parameters are as follows: sample frequency 44100 Hz, mel bands 128, hop size 347*duration to make 128 frames , pre-fft Hamming window. In addition, PCEN is compared with log mel spectrogram, however, without enough parameter adjusting, it did not show improvement than log mel [9].

Among the studies using deep learning, the task of image classification has been developed particularly, and many methods have been published in recent years. We transform the standardized mono mel feature to 3 channels of (128×128×3), so that the traditional network structures of image classification task can be applied here in this task.

3. TRAINING AND INFERENCE

3.1. Neural Network Architectures

Considering that a large amount of data contains noisy labels, we trained two models, one just used curated subset and the other used all data. The InceptionV3 structure proposed by Szegedy et al [10] was used to train the curated data and the best leadboard score can be 0.713. We trained all data using a cnn model, As Table 1 describes, the input size is (128×128×3). The model has 4 convolutional block, while each convolutional block consists of two convolutional layers followed by a average pooling layer. After each convolution, which use the rectifier (ReLU) activation function, batch normalization [11] is applied as a form of regularization. After the convolutional blocks, each channel is averaged to a scalar value. Finally, a linear layer is used to generate the predictions.

While training these two models, the training set was split into five cross-validation folds, 5 models are generated and are used for the final prediction of test set with model ensemble. The binary cross-entropy function was used as the training loss and Adam was used as the gradient descent algorithm. CosineAnnealing was used for learning rate. Label-weighted label-ranking average precision (lwLRAP) was used for evaluating the performance of the designed systems.

Table1: Description of the neural network architecture. The first parameter in each line are the kernel size filters ,"BN" refers to batch normalization .

Input(128×128×3)
3×3Conv2d(pad-1,stride-1)-64-BN-ReLU 3×3Conv2d(pad-1,stride-1)-64-BN-ReLU
2×2Avg-Pooling
3×3Conv2d(pad-1,stride-1)-128-BN-ReLU 3×3Conv2d(pad-1,stride-1)-128-BN-ReLU
2×2Avg-Pooling
3×3Conv2d(pad-1,stride-1)-256-BN-ReLU 3×3Conv2d(pad-1,stride-1)-256-BN-ReLU
2×2Avg-Pooling
3×3Conv2d(pad-1,stride-1)-512-BN-ReLU 3×3Conv2d(pad-1,stride-1)-512-BN-ReLU
2×2Avg-Pooling
Dense128(output:80)

3.2. Learning from Noisy Labels

To effectively use the data set with noisy labels, several techniques are used and compared. The first technique was proposed in [12]. It dynamically update the targets based on the current state of the model so to bootstrapped target tensor use predicted class probability directly to generate regression targets. The second technique was proposed in [13], in this paper, batch-wise loss mask is used to eliminate the several data with largest error to gradient calculation. In addition, data which will be eliminated in cross-entropy is chosen for every batch, and it is expected to allow to find noisy data gradually. The third technique we used is the tensorflow loss function: weighted_cross_entropy_with_logits, The function calculate the sigmoid cross entropy function with weight. Although we're not sure that's the right way to use it to reduce negative effect of noisy labels, but it improves the public leadboard score by 0.009. The final technique is MixMatch, that work by guessing low-entropy labels for data-augmented unlabeled examples and mixing labeled and unlabeled data using MixUp. On CIFAR-10 with 250 labels, it reduce error rate by a factor of 4 (from 38% to 11%) and by a factor of 2 on STL-10.

We modified all the softmax_cross_entropy loss of multiclass used in the above methods to sigmoid_cross-entropy loss to suitable for multilabel classification, and then trained the CNN model in Tabel 1 to observe whether the model's performance could be improved. Finally, tuning the parameters of the corresponding methods to find which could help improving public score. The effects of different techniques on the model are shown in Table 2.

Table 2: Description of the effects of noise data techniques on the model in Table 1. Test score is the best score for different parameters.

Method	Parameter	Test Score
BCE	-	0.669
First [12]	0.3	0.658
Second [13]	0.8	0.666
Third	0.7	0.678
MixMatch	-	0.602

3.3. Data Augmentation

Data augmentation is useful approach to reduce overfitting during training. We used a method called mixup [14] as the main data augmentation method. Mixup operates on a batch of train data by randomly mixing the inputs and their associated target values. Consider a pair of inputs, x_1 and x_2, and their one-hot-encoded target values, y_1 and y_2. To mix these, a parameter, $\lambda \in (0,1)$, is used to create convex combinations.

$$x = \lambda x_1 + (1 - \lambda)x_2. \quad (1)$$

$$y = \lambda y_1 + (1 - \lambda)y_2. \quad (2)$$

The output, x, y, is then used as the training example rather than the original examples. In our system, the parameter λ was a random variable from the Beta distribution B(1.0, 1.0), and a different value was used for each mixing pair. The test score of model in Table 1 was improved by 0.01 with Mixup.

3.4. Ensembling

Model ensemble is a very effective technique to increase accuracy on machine learning tasks. A good ensemble contains high performing models which are less correlated. Geometric averaging method is used here, with predictions of Inception model and cnn-8 model combined with different weights. The best public leaderboard score is 0.737.

4. RESULTS

For the curated data and all data including noisy labels, we tried several models (InceptionV3, ResNet18/34/50, CNN10 et al). We list the best models and corresponding lwlrap, leadboard score as shown in Table 3. we took the ensemble of curated LB score of 0.713 and all data LB score of 0.678, The public leaderboard score is 0.734.

Table 3: Training set results.

data	Model	lwlrap	LB score
curated	Inception V3	0.78	0.713
curated+noisy	Cnn-8	0.61	0.678
	Inception V3	0.56	0.691

5. CONCLUSION

This report described a system used to participate in Task 2 of the DCASE 2019 challenge. Many techniques were involved in feature extraction, including silence removal, computing log-mel spectrograms, mono spectrogram feature to 3 channels. We used two neural network models and combined their predictions using geometric mean. To use the data set with noisy labels well, weighted_cross_entropy_with_logits was used. To reduce overfitting, we used a data augmentation technique called mixup. In the end, the system achieved a leaderboard score of 0.737, and for the limit of kaggle kernel time, the best leaderboard is 0.734.

6. ACKNOWLEDGMENT

Thanks to Daisuke Niizumi and Dmitriy Danevskiy, who shared lots of useful tips on Kaggle. Thanks to mhiro2, who gave a concise and effective kernel as a framework of this task.

7. REFERENCES

[1] D.Giannoulis,E.Benetos,D.Stowell,M.Rossignol,M.La- grange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *2013 IEEE Workshop Appl. Signal Process. Audio, Acoustics (WASPAA)*, New Paltz, NY, 2013, pp. 1–4.

[2] 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

[3] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. *Audio set: an ontology and human-labeled dataset for audio events.* In Proc. IEEE ICASSP 2017. New Orleans, LA, 2017.

[4] Y.LeCun,Y.Bengio,andG.Hinton,"Deeplearning,"*Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] S. Mun, S. Park, D. K. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane," in *Proc. Detection Classification Acoust. Scenes Events 2017 Workshop (DCASE)*, Munich, Germany, Sep. 2017, pp. 93–102.

[6] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks," in *Proc. Detection Classification Acoust. Scenes Events 2017 Workshop (DCASE)*, Munich, Germany, Sep. 2017, pp. 80–84.

[7] Y.Xu,Q.Kong,W.Wang,andM.D.Plumbley,"Large-scale weakly supervised audio classification using gated convolutional neural network," *arXiv preprint arXiv:1710.00343*, 2017.

[8] Berthelot D , Carlini N , Goodfellow I , et al. MixMatch: A Holistic Approach to Semi-Supervised Learning[J]. 2019.

[9] Lostanlen V , Salamon J , Cartwright M , et al. Per-Channel Energy Normalization: Why and How[J]. IEEE Signal Processing Letters, 2018, 26(1):39-43.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[11] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, 2015, pp. 448–456.

[12] Reed S , Lee H , Anguelov D , et al. Training Deep Neural Networks on Noisy Labels with Bootstrapping[J]. Computer Science, 2014.

[13] Y. Jeong and H. Lim,"Audio tagging system for dcase 2018: focusing on label noise, data augmention and its efficient learning,"proceedings of DCASE 2018,2018.

[14] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *6th Int. Conf. Learn. Repr. (ICLR)*, Vancouver, Canada, 2015.