AUDIO SCENE CALSSIFICATION BASED ON DEEPER CNN AND MIXED MONO CHANNEL FEATURE

Technical Report

Nai ZHOU, Yanfang LIU, Qingkai WEI

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

ABSTRACT

This technical report describes Kuaiyu team's submissions for Task 1 - Subtask A (Acoustic Scene Classification, ASC) of the DCASE-2019 challenge[1][1]. Refering the results of DCASE 2018[1], a convolution neural network and log-mel spectrogram generated from mono audio are used, log-mel specture is converted into multiple channels spectrogram and as a input to 8 convolutional layer neural networks. The result of our experiments is a classification system that achieves classification accuracies of around 75.5% on the public Kaggle-Leaderboard[2].

Index Terms— DCASE 2019, acoustic scene classification, three, convolutional neural network, mixup

1. INTRODUCTION

Audio information obtained by auditory sense plays a very important role for human behavior. Human ears are trained by everyday life and can grasp surrounding circumstances even from fine sounds. For example, if you hear the birds singing in a quiet environment, the place is outside, where there are many foods, you can see that there are easy-to-stop trees of birds. If you have more knowledge you can also distinguish seasons and time from bird types. If the computer can automatically recognize the acoustic scene at the same level as a human being, it can be applied to various fields. For example, autonomous robots are currently mostly those that recognize information obtained from cameras and people's words. In addition to these, if it is possible to recognize the acoustic scene, it can be considered that it is possible to change the behavior of the robot and to give variations to the dialogue. However, the environmental sound continues to change over time, and the same sound will not offen occur again. Humans can respond flexibly to trivial changes in sound depending on experience, but it is extremely difficult to automate with computers. The acoustic scene classification (ASC) is one of the research subjects which is currently actively undertaken and DCASE hosted by IEEE Audio and Acoustic Signal Processing (AASP) is one of the large tasks of ASC research.

Deep learning based approaches are concentrated in this study since deep neural networks have become a state-of-the-art system for ASC, thanks to the recent advances in deep learning research, especially in the fields of image recognition and speech recognition. In the field of acoustic signal processing, CNN-based models have shown a great improvement in performance as opposed to most traditional classifiers. Several popular CNN structures have been proposed sequentially, such as AlexNet[2], VGG[3], GoogLeNet(Inception)[4], and ResNet[5]. CNNs are normally used to extract 'deep features' from the spectrograms of segmented audio waves for ASC task.

In this technical report, we tried to use mixed mono channel log-mel spectrogram features as input to train CNN network. The following section explains the details of the system architecture we proposed, experimental results, and conclusion.

2. SYSTEM ARCHITECTURE

This section describes the audio preprocessing method used in this experiment. It also describes the architecture of the neural network.

2.1 Audio Preprocessing

We use log-mel spectrogram as audio feature. Log-mel spectrogram is used by most of teams of DCASE 2018 and is considered to be effective for acoustic scene classification. The DCASE 2019 data recorder using 48kHz sampling rate and 24 bit resolution. The original recordings were split into segments with a length of 10 seconds that are provided in individual files. Available information about the recordings include the following: acoustic scene class, city, and recording location. We use the original sample for Task A and down-sample to 44.1kHz for Task B and Task C , the window size is 2,056 samples, and the hop size is 1,024 samples.Finally, Log-mel spectrogram is obtained by applying Mel filter bank, The number of bandpass filters was 128.

2.1.1 Binaural audio feature

[1]http://dcase.community/challenge2019/

[2]https://www.kaggle.com/c/dcase2019-task1a-leaderboard/leaderboard

For Task A in the DSACE 2019, the dataset is recorded using binaural microphone. From the experimental results of Han and Yuma et al.[6][7]. It turns out that using dual channel audio leaves better results than using mono audio. So we try to use binaural audio data of dual channels (Left and Right). it is presumed that it is a factor that holds more spatial information than mono audio. By calculating the log-mel spectrogram with the parameters of section 2.1, we can obtain data of (468×128×channel) shape from one audio clip's each channel. Inspired by this, we attempt to combine the features of three identical mono channels into a three-channel feature. And, in our own experiment, the latter one attain higher score than the former one. At last, for all Subtask of Task1 in the DSACE 2019, we decide to adopt mix mono channel log-mel spectrogram.

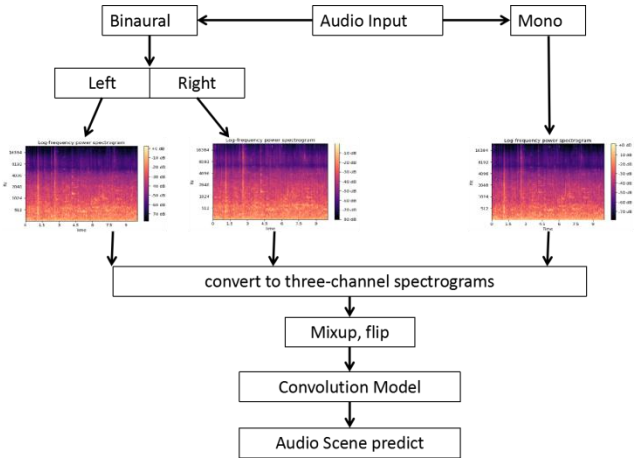

Figure 1: Architecture of the proposed system.The three-channel spectrograms are generated from one piece of audio data.We learn Network using these, and finally Ensemble learning.

2.2.2 Feature Processing

The feature we adopted are very similar to those proposed by Yuma[7]. In Yuma's work, they proposed new features inspired by image features. Many tasks of image classification use data of three channels of RGB. Among the studies using deep learning, the task of image classification has been developed particularly, and many techniques have been published in recent years. Based on the previous techniques, we use features of 3 channels(left, right, mono) so that the method of image classification task can be applied to sound classification. For Task A, the 3 channels' features contain left, right and mono, shown as Figure 1. For Task B and Task C, the 3 channels' features contain three mono.

2.2 Network Architecture

The Network we used is a 8 layers convolution neural network, as shown in Figure 2. First, we extracte the features log-mel spectrogram from raw waveforms, and convert to mix mono channel log-mel spectrogram as the input of network. The output of the neural network is the probabilities of 10 classes, between 0 and 1, with sum as 1. As an activation function within the network we use Rectified Linear Units (ReLUs).

3. DATA SIMPLE PROCESSING

We did some simple processing on the audio data for data augmentation. The methods such as mixup[8] and data inversion, are applied to the frequency domain features, which help eliminating overfitting effectively. We also try to extract the log-mel spectrogram from audio files of different time lengths, we found different time lengths has little effect on the results.

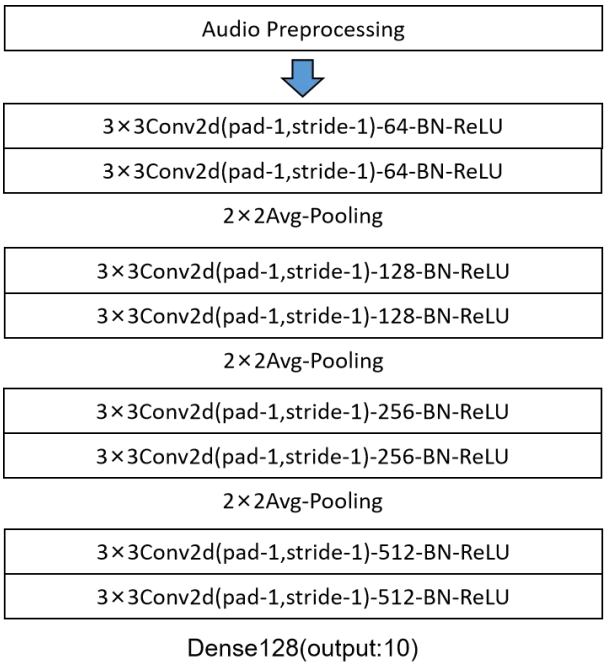

Figure 2: Convolution model called "Classifier Cnn" used in this task.

4. EXPERIMENTS

4.1 Datasets

The dataset for this task is the TAU Urban Acoustic Scenes 2019 dataset, consisting of recordings from various acoustic scenes. This dataset extends the TUT Urban Acoustic Scenes 2018 dataset with other 6 cities to a total of 12 large European cities. The dataset includes 10 scenes which are Airport, Indoor, shopping mall, Metro station, Pedestrian street, Public square, Street with medium level of traffic, ravelling by a tram, Travelling by a bus, Travelling by an underground metro, and Urban park. For each scene class, recordings were done in different locations; for each recording location there are 5-6 minutes of audio. The original recordings were split into segments with a length of 10 seconds that are provided in individual files. Available information about the recordings

include the following: acoustic scene class, city, and recording location.

For subtask A in task1, Acoustic Scene Classification is the typical acoustic scene classification task as encountered previously, where all data, both development and evaluation, are recorded with a high quality device. In this subtask there is no mismatch in recording conditions besides the natural variation of weather, people at the scene, etc, which are not under control of the recorder, but are natural manifestations of the recorded scenes.

4.2 Experiment setting

The Development Dataset setup provided by the organizer are used for the experiment. While training these models, the training set was splited into five fold cross-validation. Adaptive Moment Estimation (Adam) was used as the optimizer. Learning rate were adjusted by cosine annealing algorithm with max and min lr as 1e-3, 1e-7. The batch size differs depending on the division method.

5. RESULT

We compared the results of different networks, such as Inception, Resnet18, Rsenet34 and Classifier Cnn. We also calculated the scores of different features under the same model separately. The experimental results as shown in Table.1. Since all model uses all data of Development dataset, it describes only Leaderboard score. From Table 1, it can be seen that among the four models, Classifier Cnn got a much higher score. And for the same model, the Mix mono channel feature performed more better.

Algorithms	Channel	Accuracy (Leaderboard)
Inception V3	Mix mono channel	0.600
Resnet18	Mix mono channel	0.550
Resnet34	Mix mono channel	0.635
	Mix three channel	0.615
Classifier Cnn	Mix mono channel	**0.755**
	Mix three channel	0.716

Table.1 The experimental results.

6. CONCLUSION

In this paper, we described how to identify acoustic scenes using multiple channels Log-mel spectrogram from binaural audio and mono audio. We trained eight neural networks from the generated features and further improved accuracy by ensemble learning using these outputs. As a result, accuracy of 0.755 was obtained in the Leaderboard dataset. However, the method in this report requires training of many networks, which is not an excellent method from the viewpoint of computational resources. In future, we plan to pursue various parameter adjustment and application method of image classification method.

7. REFERENCES

[1] http://dcase.community/challenge2019/.

[2] Krizhevsky A, Sutskever I and Hinton G E, Imagenet classification with deep convolutional neural networks. Communications of the Acm, 2012. 60(2): p. 2012.

[3] Ren Z, Pandit V, Qian K, Yang Z, Zhang Z, and Schuller B. Deep sequential image features for acoustic scene classification. in The Workshop on Detection & Classification of Acoustic Scenes & Events. 2017.

[4] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, and Rabinovich A. Going deeper with convolutions. in 2015 IEEE Computer Vision and Pattern Recongition(CVPR)

[5] He K, Zhang X, Ren S and Sun J, Deep residual learning for image recognition, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[6] Y. Han, J. Park and K. Lee, “Convolutional Neural Networkswith Binaural Representations and Background Subtraction for Acoustic Scene Classification,”, IEEE AASP Challenge on DCASE 2017 technical reports, 2017.

[7] Yuma Sakashita, Masaki Aono, Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions. IEEE AASP Challenge on DCASE 2017 technical reports, 2018.

[8] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” ar-Xiv preprint ar-Xiv:1710.09412, 2017.