ACOUSTIC SCENE CLASSIFICATION BASED ON THE DATASET WITH DEEP CONVOLUTIONAL GENERATED AGAINST NETWORK

Technical Report

Fangli Ning[*1], *Shuang Duan*[1], *Pengcheng Han*[1], *Juan Wei*[2], *Zhikai Ding*[2]

[1] Northwestern Polytechnical University School of Mechanical Engineering, 127 West Youyi Road, Xi'an, 710072, China
[2] Xidian Univerisity, School of Telecommunications Engineering, South Taibai Road, Xi'an, 710071, China
ningfl@nwpu.edu.cn

ABSTRACT

As is known to us all, Convolutional Neural Networks have been the most excellent solution for image classification challenges. From the results of DCASE 2018 [1], the Convolutional Neural Network has also achieved excellent results in the classification of acoustic scenes. Therefore, our team also adopted Convolutional Neural Network for DCASE 2019 Task 1a. In order to make the audio features are exposed more, our team used multiple Mel-spectrograms to characterize the audio, trained multiple classifiers, and finally weighted the prediction results of each classifier to make results ensemble. The performance of classifier is largely limited by the quality and quantity of the data. From the results of the technical report [2], the use of GAN to augment the data set can play a vital role in the final performance, and our team also introduced Deep Convolution GANs (DCGAN) [3] to our solution to Task 1a Challenge, Our model ultimately achieved an accuracy of 0.846 on the development set and an accuracy of 0.671 on the leaderboard dataset.

Index Terms- Convolutional Neural Network, DCASE 2019, Mel-Spectrogram, data augmentation, results ensemble, DCAGN

1. INTRODUCTION

Sound and images are important media for humans to access outside information. So far, humans have achieved exciting results in the fields of image recognition, classification, and target detection, but their achievements in the field of sound classification and recognition are still not enough. The sound signals in the space of human life are still not well utilized. In some specific occasions and at certain times, the sound signal contains far more information than the image signal. For example, you can judge whether the other person is happy or sad through the voice, even if the guy keeps smiling during the communication. Although the acoustic scene classification (ASC) [4] provides several datasets, but there is still a huge gap in the number of existing image data sets. Humans still have a lot of exploration of sounds in nature and Computer Listening to do.

Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge organized by IEEE Audio and Acoustic Signal Processing (AASP) Technical Committee is one of the first large-scale challenges for ASC research. DCASE challenge has attracted a lot of sound researchers since it is launched. They provided a lot of good ideas for the acoustic scene classification and submitted very good results. [5] has achieved 81% accuracy in Task 1a of DCASE 2018, which is an exciting result. Compared to Task 1a 2018, Task 1a 2019 has more challenging, from the original 6 cities to 12 cities, followed by a linear increase in data. In theory, the increase in the amount of data is good for the training and learning of the classifier, but because of the large differences between cities, there is not much correlation between the same types of acoustic scenes in different cities. That is, the common features of the same acoustic scenes between different cities are few or difficult to find and represent, which brings great difficulties in identifying and classifying. Another great difficulty is that the official development set only contains 10 cities, while the final evaluation set contains the audio data of 12 cities. The universality of this classifier brings a lot of Challenge, because the classifier performs poorly for samples that have not appeared in the training set or that are inconsistent with the data in the training set. Therefore, when classifier developed on 10 urban data sets are applied to data samples from 12 cities, the accuracy of the classifier will be reduced, so Task 1a 2019 puts high demands on the generalization of the system.

Deep learning is based on big data. Even with the previous challenges [6-7], the amount of data has been greatly improved, but the problems to be dealt with are more and more complicated and more difficult, and the amount of data is still insufficient. This technical report will represent our contribution to the exploration, research and resolution of Task 1a 2019 challenge. This report will introduce our team's technical solutions in feature extraction, data augmentation, model building, and results ensemble.

2. SYSTEM ARCHITECTURE

This part mainly consists of audio feature extraction and the construction of the network system framework we built.

2.1. Feature extraction

The official development set contains 40 hours of audio data, which is divided into 14400 (144 per class per city) segment of 10s, the audio sampling frequency is 48KHz. Our team performs stratified sampling at a ratio of approximately 8:2 to select 11520 (116 per class per city) segment 10s audio as the training set, and the rest as the testing set.

From the technical reports submitted by the DCASE 2018 team and their excellent results, Log-Mel energy as an audio feature can reflect the acoustic scene characteristics, so our team also chose Log-Mel Energy as the audio feature. Since the Fourier transform is only suitable for stationary signals, and the general sound signals are non-stationary, we firstly perform short-time Fourier Transformation (STFT). The window size is 2048 samples, and the hop size is 1024 samples. Then add Mel filters (128) to get the Mel-spectrogram. Finally, the frequency is mapped from the linear scale to the Log-Mel scale. Because the human ear's perception of frequency is not linear, after mapping, the frequency is transformed from the linear domain to the Log-Mel scale domain, and the human ear's perception of frequency becomes linear. This transformation simulates the human ear. The perception of frequency can better reflect the audio characteristics after the transformation.

As shown in Figure 1, we obtained three Log-Mel spectrograms based on the same audio through a series of transformations. We believe that the three spectrograms can fully embody the characteristics of an audio.

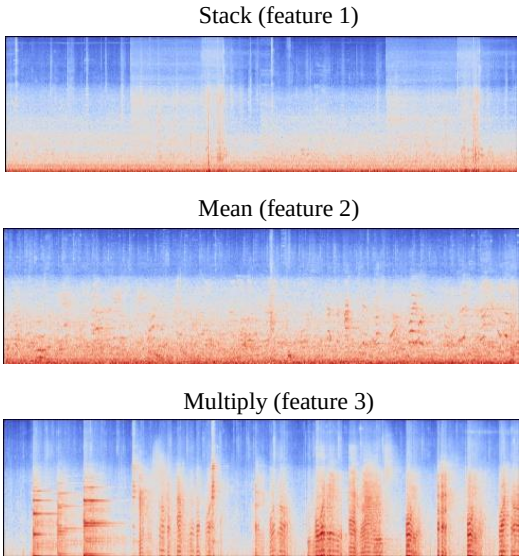

Figure 1: Stack(feature 1):stack the Left channel and the Right channel data (Left is in front of Right); Mean(feature 2): average the data of the Left and The Right; Multiply(feature 3): Multiply the data the Left and the Right.

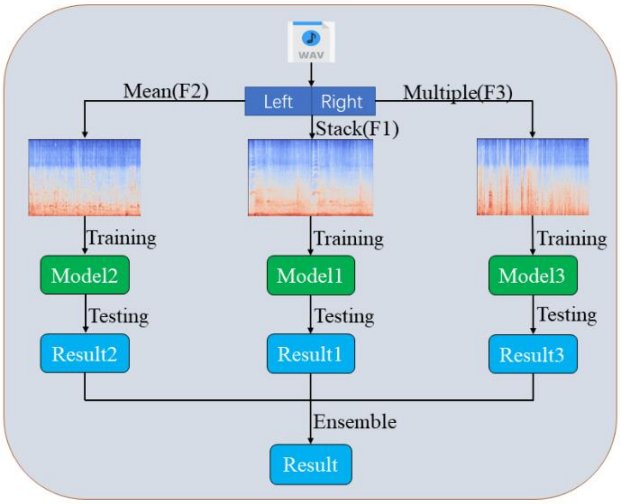

Figure 2: Extracting three features from the audios and training three classifiers, then we ensemble the predictions obtained by the three classifiers.

2.2. Network Architecture

The VGG [8] system is the runner-up in the ILSVRC [9] competition 2014, and it has become one of the most widely used systems due to its excellent performance in multiple migration tasks. Our network architecture design is based on VGG, and for the practical problems we are dealing with, our team proposed our system based on the VGG system. The system we finally submit is as described in Figure 3.

Input 469 * 128 * 3
Conv1_1 3x3 (padding-1, strides-2) - 32 - ReLU Conv1_2 3x3 (padding-1, strides-1) - 32 - ReLU Max-pooling1 2x2
Conv2_1 3x3 (padding-1, strides-1) - 64 - ReLU Conv2_2 3x3 (padding-1, strides-1) - 64 - ReLU Max-pooling2 2x2
Conv3_1 3x3 (padding-1, strides-1) - 128 - ReLU Drop-Out(0.3) Conv3_2 1x1 (padding-1, strides-1) - 128 - ReLU Max-pooling2 2x2
Conv4_1 3x3 (padding-1, strides-2) - 256 - ReLU Drop-Out(0.3) Conv4_2 1x1 (padding-0, strides-1) - 10 - ReLU Drop-Out(0.3)
Global – Average -Pooling

Figure 3: This classifier has a total of 9 layers, of which the first 8 layers are convolutional layers and the ninth layer is global average pooling.

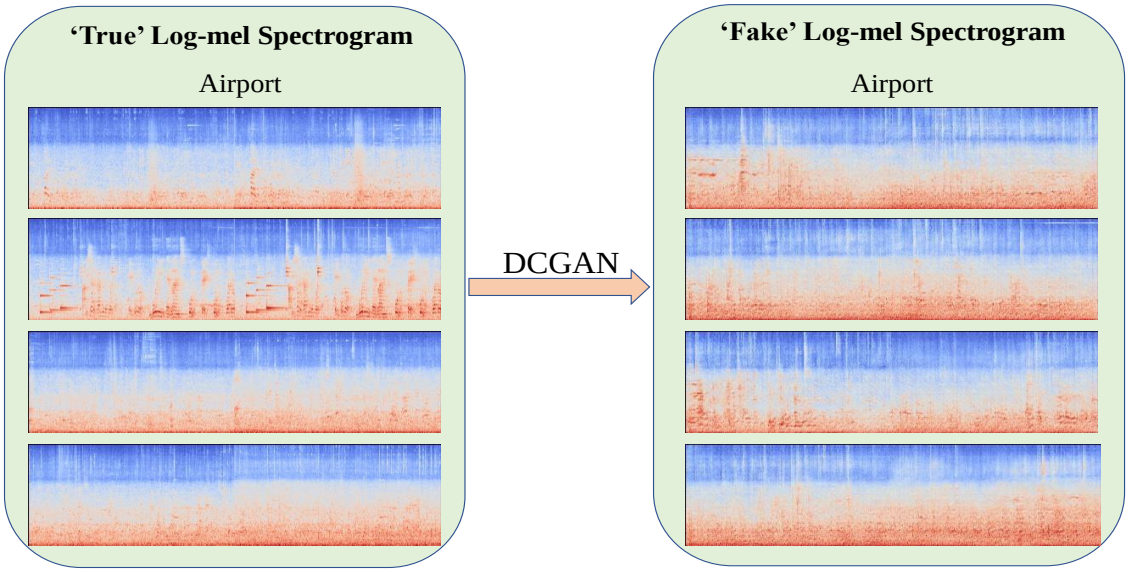

Figure 4: On the left is the Log-Mel spectrogram based on the audio, and on the right is the "Fake" Log-Mel spectrogram generated by DCGAN on the basis of the real picture.

3. DATA AUGMENTATION

In the 2017 ASC, [10] combined the Generative Adversarial Network (GAN) with the SVM Hyper-plane as a method to augment the data set and achieved very good results. From the final result, the generated "Fake" pictures as the training set has an improvement of the final prediction accuracy of the model by about 12 percentage points, which confirms the feasibility of using the generated "Fake" pictures as the training set. The "Fake" picture generated by GAN is a sample that approximates the distribution of real data. The "Fake" sample generated by it can not only expand the number of data sets, but also increase the generalization ability of the model. DCGAN is developed on the basis of GAN. They are composed of a Generator network (G) and a Discriminator network (D), but there is a big difference within the network. DCGAN is well solved the problem that the original GAN training is unstable. Compared with GAN, DCGAN has made important changes in the following aspects:

- Stride convolution is used instead of pool in Generator and Discriminator, which can make the network learn down sampling
- remove fully connected hidden layers for deeper architectures, which can greatly reduce the parameter size of the network, and is beneficial to improve the stability of the model, but also reduces the convergence speed of the model
- use batch normalization in both the Generator and the Discriminator. Accelerate the convergence of the model. It has been widely used in other network structures.
- use ReLU activation in Generator for all layers
- Using LeakyReLu activation in the Discriminator for all layers

the last two changes can prevent the model gradient disappearing and accelerate the convergence speed of the model.

As is shown in Figure 4, the "Fake" pictures we generated using the DCGAN have a little difference with the "True" pictures, but there is a certain difference in the memory. The "True" image takes up about 30KB, and the generated image is only about 15KB, which means that the generated image is less colorful than the "True" image.

4. EXPERIMENT

4.1. EXPERIMENT SETTING

We mixed the ten "Fake" classes images generated with the original ten "True" images as the final training set which consists of 27700 images. ADAM [11] algorithm was used in the process of training, with batch-size set to 64, learning rate set to 0.0003. Our model runs on NVIDIA 1080Ti and each model takes about 1.5 hours. After the model training is completed, the evaluation is performed on the divided testing set to test the model performance.

4.2. RESULTS ENSEMBLE

We let the trained three models predict the unlabeled samples on the test set separately, and then use the "weighted voting" method to ensemble the prediction results of the three models. From the previous results, we can see that Stack has the highest accuracy, followed by Mean, and the lowest is Multiple, we give it a weight of 0.45, 0.3, and 0.25, and determine the label of the sample after the calculation.

5. RESULTS

Table.1 shows the accuracy of our model on the testing set. From the last row, we can know that classifier using the Stack feature is

best, followed by the Mean and finally the Multiple. The last column of the average only contains the accuracy of the three classifiers we trained. From the last column we can know that it is more difficult for metro_station and street_pedestrian to make effective predictions than other scenes.

scenes	Baseline	Stack	Mean	Multiple	Average
Airport	0.484	0.922	0.915	0.822	0.886
Bus	0.623	0.933	0.952	0.859	0.915
Metro	0.651	0.918	0.837	0.641	0.799
Metro_station	0.545	0.801	0.741	0.578	0.706
Park	0.831	0.977	0.922	0.907	0.936
Shopping_mall	0.594	0.918	0.782	0.667	0.789
Public_square	0.407	0.890	0.852	0.563	0.768
Street_traffic	0.867	0.874	0.837	0.822	0.844
Street_pedestrian	0.609	0.778	0.752	0.659	0.730
Tram	0.640	0.782	0.807	0.759	0.783
Average	0.625	0.880	0.840	0.728	

Table 1: The accuracy of the classifiers trained by the three features on the testing set.

scenes	Stack	Stack_DCGAN
Airport	0.922	0.8333
Bus	0.933	0.9444
Metro	0.918	0.8630
Metro_station	0.801	0.7556
Park	0.977	0.9741
Shopping_mall	0.918	0.9148
Public_square	0.890	0.7593
Street_traffic	0.874	0.9296
Street_pedestrian	0.778	0.7370
Tram	0.782	0.8593
Average	0.880	0.8574

Table 2: The Stack_DCGAN represents the prediction accuracy of the classifier in which we trained the data with augmentation.

As is shown in Table 2, the average accuracy seems to have a little decrease in the data set with augmentation. We speculate that may be due to the fact that the audio of the same kind of acoustic scene has a large gap between different cities, but the dataset of the trained DCGAN is the audio of ten cities. Together, the generated image fades feature between specific cities. As a result, the accuracy of the classifier trained after data augmentation in ten cities is reduced, but this may help to improve the generalization ability of the model.

6. CONCLUTIONS

This technical report mainly describes the methods we use on the Task 1a 2019 challenge, including audio processing, feature extraction, model building, and result fusion. Although our model performed well on the testing set divided in the development set, we only achieved 0.67 on the Kaggle-leaderboard. This result challenges the generalization of our model. In the future work, we will focus on the study of the generalization of the deep learning model.

7. REFERENCES

[1] http://dcase.community/challenge2018/task-acoustic-scene-classification-results-a.
[2] Mun S, Park S, Han D K, et al. Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane[J]. Proc. DCASE, 2017: 93-97.
[3] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.
[4] http://dcase.community/challenge2019/task-acoustic-scene-classification.
[5] Sakashita Y, Aono M. Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions[J]. IEEE AASP Challenge on DCASE 2018 technical reports, 2018.
[6] http://dcase.community/challenge2016/
[7] http://dcase.community/challenge2017/
[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014
[9] http://image-net.org/challenges/LSVRC/2014/
[10] Mun S, Park S, Han D K, et al. Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane[J]. Proc. DCASE, 2017: 93-97
[11] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.