JSNU_WDXY SUBMISSION FOR DCASE-2019: ACOUSTIC SCENE CLASSIFICATION WITH CONVOLUTION NEURAL NETWORKS

Technical Report

Xinxin Ma, Mingliang Gu, Yong Ma

Jiangsu Normal University, China
School of Physics and Electronic Engineering
mxxjsnu@163.com

ABSTRACT

Acoustic Scene Classification (ASC) is the task of identifying the scene from which the audio signal is recorded. It is one of the core research problems in the field of Computational Sound Scene Analysis. Most of current best performing Acoustic Scene Classification systems utilize Mel scale spectrograms with Convolutional Neural Networks (CNNs). In this paper, we demonstrate how we applied convolutional neural network for DCASE 2019 task1, acoustic scene classification. First, we applied Mel scale spectrogram to extract acoustic features. Mel scale is a common way to suit frequency warping of human ears, with strict decreasing frequency resolution on low to high frequency range. Second, we generate Mel spectrogram from binaural audio, adaptively learn 5 Convolutional Neural Networks. The best classification result of the proposed system was71.1% for Development dataset and 73.16% for Leaderboard dataset.

Index Terms— DCASE 2019, acoustic scene classification, convolution neural networks, Mel scale spectrogram

1. INTRODUCTION

Sounds carry a great deal of information about our environments, from individual physical events to sound scenes as a whole. The problem of sensing and understanding the environment in which a sound is known as Acoustic Scene Classification, the goal is to categorize an audio recording into one of a set of (predefined) categories[1]. For example, a sound scene classification system might classify audio as one of set of categorize including home, street, and office. ASC has been applied to smartphones, tablets, robots, and cars for customized services. For example, if a car “hears” children yelling from behind a corner, it can slow down to avoid a possible accident. A smartphone could automatically change its ringtone to be most appropriate for a romantic dinner, or an evening in a noisy pub.

ASC has been a major task in the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) since 2013. In the 2013 DCASE Challenges[2], provide benchmark data for computational sound scene analysis research, including task for detection and classification of acoustic scenes and events, motivating researchers to further work in this area. Looking at the current trend of challenge submission in the ASC task, it is clear that researchers are moving traditional machine learning (Nearest Neighbor, SVM, Random Forest) to deep learning (DNNs, CNNs, LSTMs). In the DCASE 2016 ASC challenge[3], a deep CNN solution was proposed and won the rank first in the challenge task. In the 2017 ASC challenge, most of the best-performing models were based on convolutional neural networks. In the latest DCASE challenge, most of the best performing ASC systems utilize spectrograms with CNNs, or combined CNNs with other models.

Apart from, Mun et al.[4] addressed the problem of data insufficiency and proposed to use the Generative Adversarial Network (GAN) to augment training data. Han et al.[5]was focused on preprocessing of input features. Fusions of CNN model with preprocessing input features led to improved overall model performance.

An import part of ASC is to define and extract properties that characterize a certain environment – audio features. Previous work on audio features extraction mainly focused on MFCC, PLP, ZCR, Mel-spectrogram. From the submission of acoustic scene classification tasks in recent years, Mel spectrogram ha better performance. Based on the research of the above acoustic scene classification, we applied Mel spectrogram to extract features in this task, and input the extracted features into CNNs for classification.

In this report, we describe four systems for task1 (ASC) in the DCASE 2019 Challenge[6]. We provide the performance of our systems on the openly accessible DCASE 2019 datasets. In our challenge submissions, we show 8 different performance for task1a. The following sections describe the details of the applied system and the experimental results and conclusions. Section 2 introduces the system architectures. Section 3 presents some experimental results. Sections 4 concludes the papers.

2. SYSTEM ARCHITECTURE

This section introduces the applied audio preprocessing methods. It also describes the details utilized process flow and ConvNet architecture.

2.1 Feature Extraction

We use Mel-spectrogram as audio features. Mel-spectrogram is considered to be most suitable for acoustic scene classification. The original audio data is first preprocessed, and the original 44.8 kHz sampling rate of the two-channel audio signal is downsampled to a 44.1 kHz mono signal. Next, frequency and phase are analyzed the Short-time Fourier Transformation (STFT). STFT can calculate the spectrum at each time by looking at the time change by multiplying the window, the window size is 2,048 samples, and the hop size is 1,024 samples. Mel-spectrogram is obtained by applying Mel filter bank. Then, we extracted the spectrogram with 64 bin Mel-scales. And, a logarithm operation is applied to obtain on the log-Mel spectrogram. Hz is converted to mel by using the following equation.

$$\text{mel} = 2595.0log_{10}(\frac{1.0+frequencies}{700.0}) \qquad (1)$$

Mel spectrogram feature is extracted using librosa[7] toolbox.

2.3 Network Architecture

CNNs has achieved state-of-the-art performance in image classification. A CNN consists of several convolutional layers followed by fully-connected layers. Each convolutional layers consists of filters to convolve with the output from the previous convolutional layers. The filters can capture local patterns in features maps, such as edges in lowers layers and complex profiles in higher layers. Meanwhile, CNNs have been applied to many audio classification and sound event detection task using inputs such as log Mel-spectrogram.

AlexNet and VGNet models are not fundamentally different from the traditional CNN models in principle, and have been gradually applied to acoustic scene classification. However, AlexNet [8] uses a 5*5 large convolution core, which can capture the finer-grained feature changes of the most original audio data as early as possible. However, with the deepening of the network depth, the feature correlation in a larger local range will be lost, while VGG-Net [9] uses a 3*3 small convolution core to improve performance by deepening the network structure.

We followed the CNN framework proposed in [10][11] with some modification. Table 1 shows the overall architecture of the applied system.

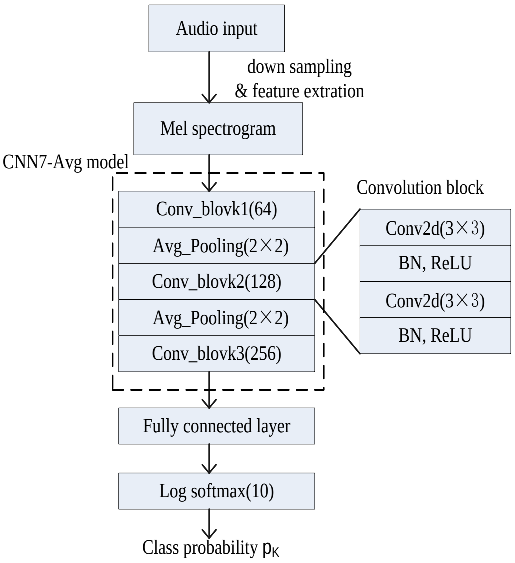


Figure 1: The overall architecture of the applied system

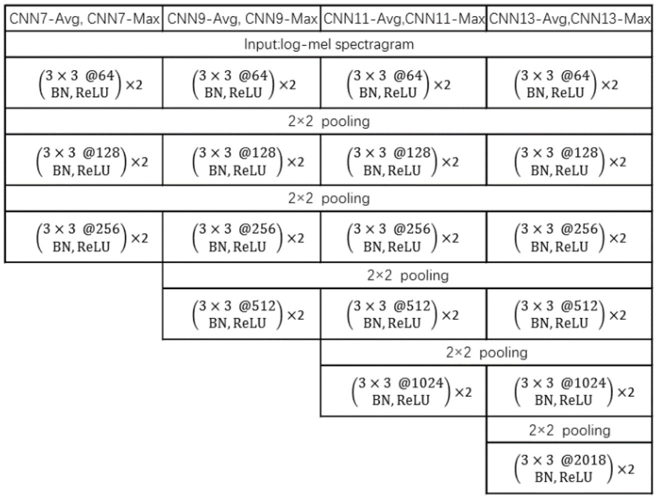
The VGG network was proposed to decompose the 5*5 kernel to a convolutional network block consisting of two cascaded convolutional layers with 3*3 kernels. So, we applied the 7-layers ,9-layers, 11-layers CNN and 13-layrs CNN in this paper with 3, 4 , 5 and 6 convolutional blocks. We apply batch normalization (BN) after each convolutional operation to speed up and stabilize the training. The ReLu(Rectified Linear Unit) function is used as a non-linearity after batch normalization. Average pooling or max pooling with a size 2*2 is applied after each convolutional block to reduce the feature map size. The last layer is composed of 10 neurons, and the activation function is classified by non-linear softmax function.

For classification task, softmax nonlinearity is applied and cross entropy (CE) loss l_{CE} is used for training the network:

$$l_{CE}(p,y) = \sum_{k=1}^{k} y_k ln p_k \qquad (1)$$

Where $y = (y_1, \dots\dots, y_k) \in \{0,1\}^K$ is the clip-level or frame-level target and K is the number of sound classes. The prediction $p = (p_1, \dots., p_k) \in [0,1]^K$ is the predicted probability of sound classes.

Figure 2 shows the details of CNN architectures used in this paper. For example, 3×3@ 64 indicates a convolutional layer with a kernel size of 3×3 and an output feature maps number of 64.



CNN7-Avg, CNN7-Max	CNN9-Avg, CNN9-Max	CNN11-Avg,CNN11-Max	CNN13-Avg,CNN13-Max
Input:log-mel spectragram			
$\begin{pmatrix}3\times3\ @64\\ \text{BN, ReLU}\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\ @64\\ \text{BN, ReLU}\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\ @64\\ \text{BN, ReLU}\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\ @64\\ \text{BN, ReLU}\end{pmatrix}\times2$
2×2 pooling			
$\begin{pmatrix}3\times3\ @128\\ \text{BN, ReLU}\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\ @128\\ \text{BN, ReLU}\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\ @128\\ \text{BN, ReLU}\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\ @128\\ \text{BN, ReLU}\end{pmatrix}\times2$
2×2 pooling			
$\begin{pmatrix}3\times3\ @256\\ \text{BN, ReLU}\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\ @256\\ \text{BN, ReLU}\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\ @256\\ \text{BN, ReLU}\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\ @256\\ \text{BN, ReLU}\end{pmatrix}\times2$
	2×2 pooling		
	$\begin{pmatrix}3\times3\ @512\\ \text{BN, ReLU}\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\ @512\\ \text{BN, ReLU}\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\ @512\\ \text{BN, ReLU}\end{pmatrix}\times2$
		2×2 pooling	
		$\begin{pmatrix}3\times3\ @1024\\ \text{BN, ReLU}\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\ @1024\\ \text{BN, ReLU}\end{pmatrix}\times2$
			2×2 pooling
			$\begin{pmatrix}3\times3\ @2018\\ \text{BN, ReLU}\end{pmatrix}\times2$

Figure 2: The details of CNN architectures used in this paper

3. EXPERIMENT

3.1 Database

To evaluate our system, we use the task1a acoustic scene classification data from the official data set of DCASE 2019[12], which is recorded in 12 European cities, such as Barcelona, Paris and Milan. For all acoustic scenes, audio is captured in many places: different streets, different parks, different shopping malls. Data collectors wear headphones made from microphones and capture multiple 2-3-minute recordings at several slightly different locations (2-4) in each location, which makes the recorded audio very similar to the sound heard by human auditory system. The TAU Urban Acoustic Scenes 2019 development dataset consists of 10 acoustic scenes: airport, bus, metro, metro_station, park, public_square, Shopping mall, pedestrian street, street_traffic, tram. Each type of scene has 1,440 two-channel audio (240 minutes of audio) lasting 10 seconds, with a sampling rate of 44.8 kHz and a quantization dimension of 24 bits. In this paper, the original 44.8 kHz sampling rate of the two-channel audio signal is downsampled to a 44.1 kHz mono signal. The data set contains a total of 40 hours of audio.

TAU Urban Acoustic Scenes 2019 Development dataset is divided into 9185 training set data and 4185 test set data according to a certain proportion, as shown in Table 1. The ASC model is trained with training set. Meanwhile, the weighted average accuracy of all sound scene categories is used as an objective criterion to evaluate the classification performance on the test set.

classes	training set	testing set
airport	911	421
bus	928	415
metro	902	433
metro_station	897	435
park	946	386
public_square	945	387
shooping_mall	896	441
street_pedestrain	924	429
street_traffic	942	402
tram	894	536
total	9185	4185

Table 1: TAU Urban Acoustic Scenes 2019 task1a development dataset

3.2 Experiment Settings

When using the Librosa tool library for Mel spectrogram feature extraction, the feature parameters are set as follows: sampling rate 44.1 kHz, using Hamming window, window size 2048 samples, window shift size 1024 samples, 64 mel bins, the cutoff frequency is between 20Hz and 16KHz.

The experiment was carried out using data setting provided by the organizer. Network training was performed by optimizing the categorical cross-entropy and Adam. The Adam optimizer with a learning rate of 0.0001 and the learning rate is reduced by multiplying 0.9 after every 200 iterations training. Also, a batch size 32 is used.

3.3 Experiment Results

Table 2 shows the classification accuracy of Mel-spectrogram acoustic characteristics in DCASE 2019 ASC baseline system and the different depth convolution neural networks we applied.

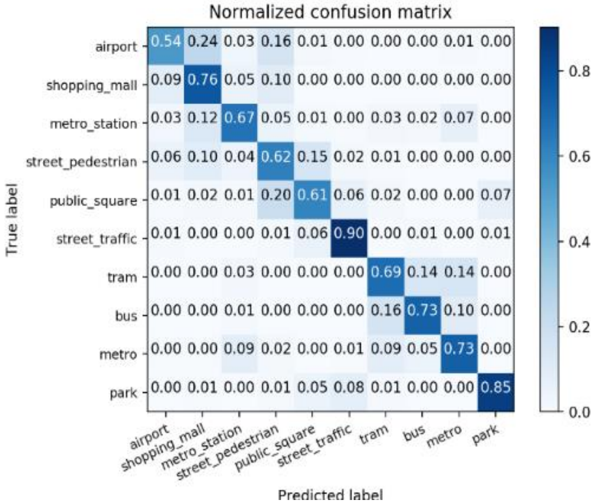
models	accuracy
Baseline	61.8%
CNN7-Avg	71.1%
CNN7-Max	69.5%
CNN9-Avg	69.9%
CNN9-Max	67.5%
CNN11-Avg	67.4%
CNN11-Max	65.5%
CNN13-Avg	60.4%
CNN13-Max	65.5%

Table 2: The accuracy of the DCASE 2019 baseline system and the different depth convolution neural networks we applied.

From the experimental results in Table 2, it can be seen that:

(1) The convolutional neural network model used in this paper is generally better than the DCASE 2019 baseline system. The baseline system of the DCASE 2019 ASC uses a 7×7 convolution kernel and a two-layer ConV2D network structure. This paper uses a 3×3 small convolution kernel based on VGGNet and constructs 4 different depth models. It can be found from the comparison of experimental results that the network structure of this paper can better implement scene classification.

(2) Comparing the different convolutional layers and different pooling types used in this paper, it can be found that CNN7-Avg classification performance is better than CNN9, CNN11 and CNN13, with the best correct rate of 71.1%. Comparing the experimental results between four sets of convolutional network models with different depths, it can be found that merely increasing the network depth does not improve the system performance.



	airport	shopping_mall	metro_station	street_pedestrian	public_square	street_traffic	tram	bus	metro	park
airport	0.54	0.24	0.03	0.16	0.01	0.00	0.00	0.00	0.01	0.00
shopping_mall	0.09	0.76	0.05	0.10	0.00	0.00	0.00	0.00	0.00	0.00
metro_station	0.03	0.12	0.67	0.05	0.01	0.00	0.03	0.02	0.07	0.00
street_pedestrian	0.06	0.10	0.04	0.62	0.15	0.02	0.01	0.00	0.00	0.00
public_square	0.01	0.02	0.01	0.20	0.61	0.06	0.02	0.00	0.00	0.07
street_traffic	0.01	0.00	0.00	0.01	0.06	0.90	0.00	0.01	0.00	0.01
tram	0.00	0.00	0.03	0.00	0.00	0.00	0.69	0.14	0.14	0.00
bus	0.00	0.00	0.01	0.00	0.00	0.00	0.16	0.73	0.10	0.00
metro	0.00	0.00	0.09	0.02	0.00	0.01	0.09	0.05	0.73	0.00
park	0.00	0.01	0.00	0.01	0.05	0.08	0.01	0.00	0.00	0.85

Figure3: Confusion matrix based on CNN7-Avg classification results.

In addition, in order to better analyze the classification accuracy rate of different scenes, the confusion matrix is utilized in the model evaluation. The confusion matrix is an N×N matrix, N represents the number of categories of the overall data, one column of the matrix represents the category of the model prediction, and one row of the matrix represents the real category to which the sample belongs. In view of the above experimental results, we chose to plot the confusion matrix of the CNN7-Avg network with the best classification accuracy. As shown in Figure 3, most acoustic scene categories are better distinguished, and confusion mainly occurs in airport and public place scene categories.

4. CONCLUSION

In order to improve the classification performance, a new convolution neural network is constructed and applied to acoustic scene classification. The network is based on VGGNet, a deep convolution neural network. By taking advantage of the advantages of the small convolution core of 3*3 and the small pool core of 2*2 of the network, the problem of feature correlation in a large part of the lost audio signal caused by the deepening of the network layer is avoided to a certain extent. The experimental results show that the network structure of CNN7-Avg has the best classification performance. Compared with the baseline system of DCASE 2019ASC, CNN7-Avg has better classification accuracy.

5. REFERENCES

[1] Virtanen T , Plumbley M D , Ellis D,"Computational Analysis of Sound Scenes and Events || Approaches to Complex Sound Scene Analysis," 2018.

[2] http://dcase.community/

[3] Mesaros A , Heittola T , Benetos E , et al, "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(2):379-393.

[4] S. Mun, S. Park, D. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane," Tech.Rep, DCASE2017 Challenge, September 2017.

[5] Y. Han and J. Park," Convolutional neural network with binaural representations and background subtraction for acoustic scene classification," In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop, November 2017, 46-50.

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

[7] B. McFee, C. Raffel, D. Liang, D. P. Ellis, et al.," librosa: Audio and music signal analysis in python." In Proceedings of the 14th python in science conference, 2015, pp.18-25.

[8] Simonyan K , Zisserman A, " Very Deep Convolutional Networks for Large-Scale Image Recognition. Computer Science," 2014.

[9] Krizhevsky A , Sutskever I , Hinton G, " ImageNet Classification with Deep Convolutional Neural Networks," NIPS. Curran Associates Inc. 2012.

[10] Kong, Qiuqiang , et al. "DCASE 2018 CHALLENGE SURREY CROSS-TASK CONVOLUTIONAL NEURAL NETWORK BASELINE." (2018).

[11] Kong, Qiuqiang , et al. "Cross-task learning for audio tagging, sound event detection spatial localization: DCASE 2019 baseline systems." (2019).

[12] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "A multi-device dataset for urban acoustic scene classification," In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), 9–13. November 2018.