ACOUSTIC SCENE CLASSIFICATION BASED ON CNN SYSTEM

Technical Report

Zhuhe Wang, Jingkai Ma, Chunyang Li
Beijing Technology and Business University
Noise and Vibration Laboratory
No.11 Fucheng Street, Beijing, China
btbuwzh119@126.com, 929542133@qq.com, chuny6896@163.com

ABSTRACT

In this study, we present a solution for the acoustic scene classification task1A in the DCASE 2019 Challenge. Our model uses a convolutional neural network and makes some improvements on the basis of CNN. Then we extract the MFCC (Mel frequency cepstral coefficient) feature from the official audio file and recreate the data set. Use this as an input to the neural network. Finally, comparing our model to the performance of the baseline system, the results were 12% more accurate than the baseline system.

Index Terms— Acoustic Scene Classification, Mel frequency cepstral coefficient, Convolutional Neural Network.

1. INTRODUCTION

In recent years, great progress has been made in audio feature extraction, and different types of time-frequency transformed images are applied for feature extraction, such as spectrum based on short-time Fourier transform [1], scalar spectrum, log- Mel spectrum [2] and MFCC spectrum [3] [4]. MFCC is widely used in audio feature extraction. A segment of speech is divided into many frames. Each frame of speech corresponds to a spectrum (calculated by short-time FFT), and the spectrum represents the relationship between frequency and energy. Among the many methods of audio feature extraction, the Mel frequency cepstral coefficient shows a unique advantage, so we perform MFCC feature extraction on the TAU Urban Acoustic Scenes 2019 dataset, extracting the discerning components of the audio signal to train as a feature.

In actual use, there are three types of spectrograms, namely linear amplitude spectrum, logarithmic amplitude spectrum, and self-power spectrum (the amplitudes of each spectral line in the logarithmic amplitude spectrum are logarithmically calculated, so the unit of the ordinate is dB (Decibel), the purpose of this transformation is to pull those components with lower amplitudes higher relative to the high amplitude components in order to observe the periodic signals that are masked in low amplitude noise). To increase the time dimension, display a spectrum of speech, and visually see static and dynamic information, we will get a spectrogram that changes over time. The properties of phonemes can be better displayed in the spectrogram, and the sound can be better recognized by observing the formants and their transitions. By performing cepstrum analysis on the Mel spectrum, the Mel frequency cepstrum coefficients are obtained, and then the speech feature vectors are imported into the network model for training and recognition.

With the speed of computer operation, deep learning has also been greatly developed. The convolutional neural network has shown good performance in the image recognition data set MNIST. Subsequently, the convolutional neural network has been greatly developed. VGG, AlexNet, GoogleNet. Convolutional neural networks are favored by deep learning researchers because of their superior performance. Convolutional neural networks also exhibit strong performance in the field of audio recognition, such as the application of abnormal sound detection in home appliances [10]. In DCASE2018, CNN also exhibits good performance [5][6][7][8]. In our paper, the MFCC features extracted from audio files are used as the main basis for acoustic scene classification. We constructed a four-layer convolution, two-layer fully connected layer convolutional neural network as our model, and perform sound field recognition on this model.

The structure of this paper is as follows, the data preprocessing and feature extraction , and the production of data sets introduced in Chapter 2. The third chapter introduces the description of the system and the experimental process and method. The fourth chapter is the evaluation of system performance and discuss.

2. PRODUCTION OF DATA SETS

The original data set of this mission contains acoustic records of 10 scenes in 12 cities. The acoustic scenes include airport， shopping mall, metro station, pedestrian street, public square, traffic street, tram, bus, metro, park. For each scene class, the recording takes place at a different location; for each recording position, there is 5-6 minutes of audio, and the original recording is divided into segments of 10 seconds in length. The development data set includes recordings from ten cities. A training/test subset is created based on the recorded location such that the training subset contains approximately 70% of the recorded locations from each city, and the test subset contains records from other locations.

The development set contains 40 hours of data and 14400 segments (144 per city for each acoustic scene category). There are 10080 audios in the training set and 4320 in the test set. The evaluation data set contains 20 hours of audio data from 12 cities (two cities not encountered in the development set).

We created a new data set based on the original data, and performed MFCC feature extraction for each 10 second audio file. It is a cepstrum parameter extracted in the Mel scale frequency domain. The Mel scale describes the nonlinearity of the human ear frequency. Its relationship with frequency can be approximated by the following formula:

$$\text{Mel}(f) = 2595 \times \log(1 + f/700)$$

Where f is the frequency in Hz. The Fig1 below shows the relationship between Mel frequency and linear frequency:

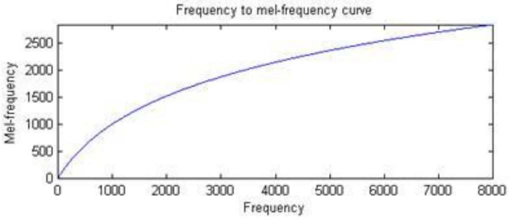

Figure1: the relationship between Mel frequency and linear frequency

The 24 subband energy features are decorrelated and reduced to 13 dimensions by DCT transform, including C0~C12. (Driving by discrete cosine transform (DCT), the role of DCT is to obtain the cepstrum of the spectrum. The low-frequency component of the cepstrum is the envelope of the spectrum, and the high-frequency component of the cepstrum is the details of the spectrum. These are all scientifically proven speech physics information for speech recognition. C0 is actually the total energy level of each sub-band. C0 is consistent with the energy coefficient. C0-C12 is the first 13 coefficients. After DCT, almost all coefficients are successively decremented to 0 after 13, so we have chosen the first 13 dimensions of the MFCC feature as the input vector.

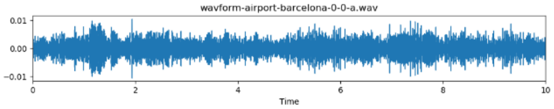

Figure2: waveform diagram of airport-barcelona-0-0-1.wav

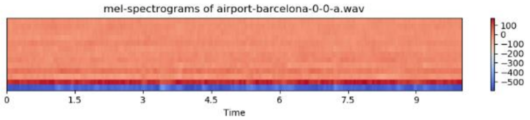

Figure3: a two-dimensional spectrum after MFCC feature extraction

Fig 2 shows a waveform diagram of an audio file, and Fig 3 shows a two-dimensional spectrum after MFCC feature extraction, the abscissa indicates time and the ordinate indicates frequency band. We resampled each audio file, and the sampling rate changed from 48khz to 22.05khz, so the time series obtained by MFCC feature extraction is 431, each audio sample corresponds to 13*431=5603 data points. The point response is a sample feature, then we relabel each sample, and 70% as a training set and 30% as a test set as a data set for our model.

3. OUR SYSTEM

After a cursory review of previously submitted articles over the past two years, it's easy to see that in-depth learning solutions are very popular among researchers. Such as CNN, RNN, DNN, GAN [11], and all of these programs have achieved good results in the challenge of the ASC tasks. However, the uncertainty or observation error of the input data accumulates over time in the RNN forward process, the RNN often becomes unstable and it is difficult to learn a reliable model. DNN also has obvious shortcomings in the application process. In contrast, the Convolutional Neural Network (CNN) uses a fixed-size kernel window, so it eliminates long-term cumulative effects. In the DCASE 2017 and 2018 leaderboard, the top three are based on the CNN system. Obviously, CNN is a powerful model of ASC tasks, and we also use the CNN model as one of the key components of our proposed framework.

We conducted several experiments on the network structure and hyperparameters of the CNN model. The results are shown in Table 1 below.

Table1: Experimental process and results

convolution layers	Convolution kernel	Pooled window	Activation function	Accuracy
4	7*17 2*2 2*2 2*2	1*34 2*2 2*2 2*2	Relu	68.33%
4	5*5 2*2 2*2 2*2	1*34 2*2 2*2 2*2	Tanh	59.28%
4	7*7 2*2 2*2 2*2	1*34 2*2 2*2 2*2	Relu	70.69%
4	13*34 2*2 2*2 2*2	1*34 2*2 2*2 2*2	Relu	73.50%
2	13*13 13*13	13*13 1*17	Relu	71.67%

Our model uses a 4-layer convolutional neural network model, and its network architecture is shown in Figure 4. the first-layer convolution kernel is 13*34, and the remaining three-layer convolution kernels are 2*2. The first layer of pooling window is 1*34, the step size is 1*34, and the other three layers of pooling windows are 2*2, and the step size is 2*2. The first layer of the fully connected layer uses 4096 nodes, and the second layer uses 10 nodes. We use the RELU activation function and the Adam optimizer. In the end we used 20,000 steps to iterate, the batch number is 100. The experimental results show that the accuracy of our model after testing on the development data set reached 73.5%.

Input 13*431
13*34 Conv(padding,stride[1,1])-32-BN-Relu
1*34 MaxPolling(padding,stride[1,34])
2*2 Conv(padding,stride[1,1])-64-BN-Relu
2*2 MaxPolling(padding,stride[2,2])
2*2 Conv(padding,stride[1,1])-128-BN-Relu
2*2 MaxPolling(padding,stride[2,2])
2*2 Conv(padding,stride[1,1])-256-BN-Relu
2*2 MaxPolling(padding,stride[2,2])
Fully connected1-4094-Relu
Dropout(0.3)
Fully connected2-10-Relu

Figure4: Network architecture

In the field of machine learning, the confusion matrix is also called the probability table or the error matrix. It is a specific matrix used to visualize the performance of an algorithm. Each column represents a predicted value, and each row represents the actual category. In the confusion matrix, all correct predictions are on the diagonal, so it is easy and intuitive to see from the confusion matrix where there are errors because they are outside the diagonal. The confusion matrix obtained by developing dataset in our model training is shown in Figure 5.

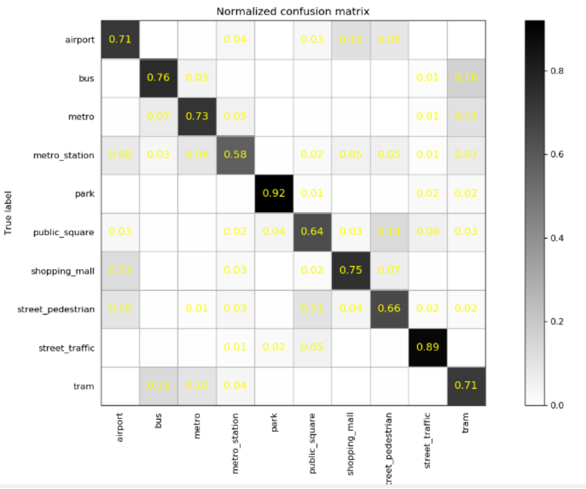

Figure 5: confusion matrix

The diagonal of Figure 5 description the test accuracy of 10 acoustic scenes. In turn, airport71%, bus76%, metro73%, metro_station58%, park92%, public_square64%, shopping_mall-76%, street_pedestrian66%, street_traffic89%, tram71%. The overall prediction accuracy of the model is 73.5%

4. RESULTS AND DISCUSSION

In this report, we present an effective solution for the Acoustic Scene Classification task 1a in the DCASE 2019 challenge, which increases the accuracy of the Acoustic Scene Classification to 73.5%. We use the MFCC feature of the audio as a sample feature and train based on CNN,and we got good results. In addition, there is still room for improvement in the insufficient of our current work. For the data processing part, data enhancement and other technologies can also be adopted. For the network part, a network with better performance can also be selected to achieve better results.

5. REFERENCES

[1] Yingjie li,’’ Research on Feature Extraction of Audio Event Detection Based on Spectrogram’’, Beijing University of Posts and Telecommunications

[2] Yuma Sakashita and Masaki Aono,’’ Acoustic Scene Classification by Ensemble of Spectrograms Based on Adaptive Temporal Divisions’’ IEEE AASP Challenge on DCASE 2018 technical reports, 2018.

[3] Juergen Tchorz’ ‘Combination of Amplitude Modulation Spectrogram Features and MFCCs for Acoustic Scene Classification’’ IEEE AASP Challenge on DCASE 2018 technical reports, 2018.

[4] Zhitong Li, Liqiang Zhang, Shixuan Du and Wei Liu.’’ Acoustic Scene Classification Based on Binaural Deep Scattering Spectra with CNN and LSTM’’ IEEE AASP Challenge on DCASE 2018 technical reports, 2018.

[5] SangwonLee, Seungtae Kang, Gil-Jin Jang.’’ CNNBASEDSYSTEMFORACOUSTICSCENECLASSIFI CATION Technical Report’’ IEEE AASP Challenge on DCASE 2017 technical reports, 2018.

[6] Zhang Liwen Han Jiqing.’’ ACOUSTIC SCENE CLASSIFICATION USING MULTI-LAYERED TEMPORAL POOLING BASED ON DEEP CONVOLUTIONAL NEURAL NETWORK’’ IEEE AASP Challenge on DCASE 2017 technical reports, 2018.

[7] W. Zheng, J. Yi, X. Xing, X. Liu and S. Peng, “Acoustic Scene Classification Using Deep Convolutional Neural Network and Multiple Spectrograms Fusion”, IEEE AASP Challenge on DCASE 2017 technical reports, 2017.

[8] Zhao Ren1, Qiuqiang Kong2, Kun Qian1, Mark D. Plumbley2, Bj¨orn W. Schuller1,3,’’ ATTENTION-BASEDCONVOLUTIONALNEURALNETWORKSFOR ACOUSTICSCENECLASSIFICATION Technical Report’’ IEEE AASP Challenge on DCASE 2017 technical reports, 2018.

[9] http://dcase.community/ challenge2018/.

[10] Jiang Y, Li C Y, Li N, Feng T, and Liu M L. HAASD: A dataset of Household Appliances Abnormal Sound Detection [C]// 2018 International Conference on Computer Science and Artificial Intelligence. (Shenzhen, China, December 8-10, 2018) ACM New York, NY, 2018.6

[11] Seongkyu Mun and Sangwook Park,” Generative Adversarial Network Based Acoustic Scene Training Set Augmentation and Selection Using SVM Hyper-Plane” IEEE AASP Challenge on DCASE 2018 technical reports, 2017.