ACOUSTIC SCENE CLASSIFICATION USING ATTENTION-BASED CONVOLUTIONAL NEURAL NETWORK

Technical Report

Han Liang

Huazhong University of Science and Technology
Wuhan National Laboratory for Optoelectronics
Wuhan, China
hanliang@hust.edu.cn

Yaxiong Ma

Huazhong University of Science and Technology
Wuhan National Laboratory for Optoelectronics
Wuhan, China
marathon@hust.edu.cn

ABSTRACT

This technical report describes the Task 1 - Subtask A (Acoustic Scene Classification, ASC) of the DCASE 2019 challenge whose goal is to classify a test audio recording into one of the predefined classes that characterizes the environment. We detemine to use mel-spectrogram as audio feature and deep convolutional neural networks (CNNs) as classifier to classify acoustic scenes. In our method, spectrogram of every audio clip is divided in two ways. In addition, we introduce attention mechanism to further improve the performance. Experimental results illustrate that our best model can achieve classification accuracy of around 70.7% for Development dataset, which is superior to the baseline system with the accuracy of 62.5%.

Index Terms— Acoustic scene classification, convolutional neural network, spectrogram division, attention mechanism

1. INTRODUCTION

Acoustic scene classification which is one of the research subjects in the DCASE 2019 Challenge aims to recognize surrounding environment using acoustic signals. It has been used in various fields, such as surveillance, autonomous driving and multimedia retrieval. At present, ASC has been attracting the attention of a large number of scholars in this area. Many researchers have proposed various acoustic scene classification systems to recognize those scenes. Nowadays, many features and methods have been applied to classify acoustic scenes. The explored acoustic features include Mel-frequency cepstral coefficients [1, 2], mel-secptrograms[3, 4, 5], Constant-Q transformed spectrograms[3], and some other low-level features[6]. The popular methods include Support Vector Machine(SVM)[6, 7], Hidden Markov Model(HMM)[8], Guassian Mixture Model(GMM)[2]. With the rapid development of deep learning techniques, neural networks quickly become the mainstream solution for ASC. Especially, convolutional neural networks (CNNs)[1, 3, 9]are successfully applied to classify acoustic scenes and have achieved the state-of-the-art performance.

In this report, we present a new acoustic scene classification method. We first divide the spectrogram of every audio clip in two ways: overlap division and non-overlap division. In addition, we introduce attention mechanism to further improve the performance.

The rest of the report is organized as follows. Section 2 describes the audio processing method, spectrogram division methods and the proposed attention mechanism. Section 3 shows the experiments performed to prove the efficiency of the proposed method. Finally, we conclude the work in Section 4.

2. SYSTEM ARCHITECTURE

The audio processing method and spectrogram division method used in our experiments will be described in this section. And in this section, we will describe the attention mechanism and the architecture of the proposed network.

2.1. Audio processing

Log mel-spectrograms are employed as audio features used in our experiments. The parameters are as followes: we extract log mel-spectrograms using 2048-point short time Fourier transform(FTFT) on 40ms Hamming windowed frames with 20ms overlap. The number of bandpass filters is 64. Finally the mel-spectrograms are transformed to the logarithmic scale. In this way, the spectrogram of every audio clip is with the shape of (64,500).

2.2. Spectrogram division

We have suggested two ways to divide the spectrogram. The first one is non-overlap division, in which the audio is divided into 5 segments with the length of 2 seconds. The second one is overlap division, in which the audio is divided into 9 segments with half overlap and each segment has the same length as method one. Experimental results illustrate that the model with overlap division is better than that with non-overlap division. The spectrogram division methods used in this technical report and the overall acoustic scene classification framework can be seen in Figure 1.

2.3. Attention mechanism

Recently, many attention-based deep neural networks have been proposed for acoustic scene classification. Attention pooling mechanism proposed in[10]can adaptively learn the contributions of the time-frequency units. Self-attention mechanism is used for modeling relationship between different positions of a spectrogram[5]. In this report, we introduce the frequency attention networks, illustrated in Figure 2. We will first acquire the frequency statistic of mel-spectrogram feature. Then, we feed the feature through a single linear layer and then a sigmoid activation function to produce

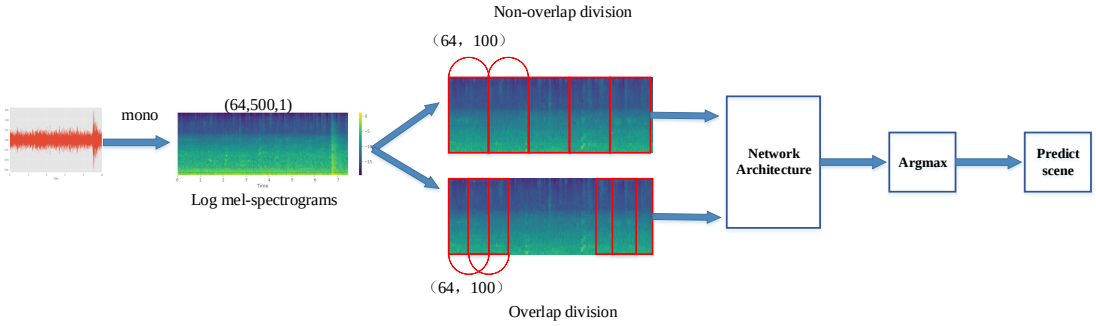

Figure 1: Acoustic Scene Classification Framework

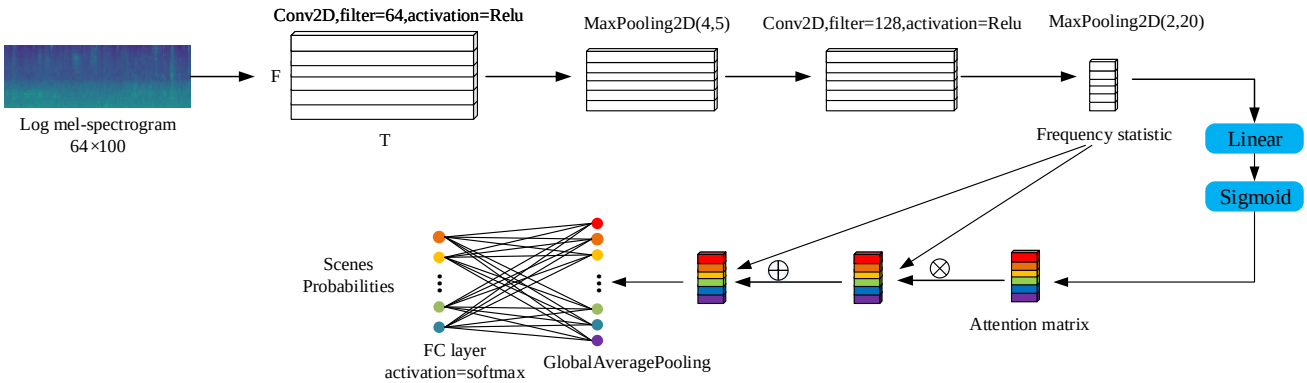

Figure 2: Structure of Attention Network.

the frequency attention distribution of the feature. Finally, we multiply frequency statistic by the attention distribution to emphasize informative frequency signal and suppress less useful ones.

2.4. Network architecture

Our system is based on the the baseline system architecture[11] depicted in Figure 3. The network contains 2 CNN blocks. The second block is frequency statistic pooling to compress the frame level layer to one statistic layer, followed by attention block. The last layer of the network is a Dense layer with 10 nodes and Softmax activation function.

3. EXPERIMENTS AND RESULTS

Experiments and results are reported in this section, including the dataset and details about the important parameters used in our model.

3.1. Datasets

Dataset for this task is the TAU Urban Acoustic Scenes 2019 dataset, consisting of recordings from various acoustic scenes. This dataset contains recordings of 10 ten acoustic scenes, recorded in twelve large European cities. Acoustic scenes included are: airport, shopping mall(indoor), metro station, pedestrian street, public square, street(medium level of traffic), traveling by tram, traveling by bus, metro(underground), urban park. TUT Urban Acoustic Scenes 2019 development dataset contains 1440 segments for each acoustic scene(240 minutes of audio), of which the number is increased compared with DCASE 2018.

3.2. Experimental setup

We implement our model in keras with Tensorflow backend and experiments are performed on a single GPU having 12GB RAM. We train models 10 times for 200 epochs and report the average best accuracy. We use Adam as optimizer with the learning rate of 0.001.

3.3. Results

Table 1 shows the classification results of our proposed method with attention and without attention on the development set. From the experimental results, we also find that the performance of overlap division method is better than non-overlap division mnethod. So we only show the performance of network model using overlap division method. Compared with the baseline system, our best model achieves a relative improvement of more than 8%. And the confusion matrix of the best model is shown in table 2.

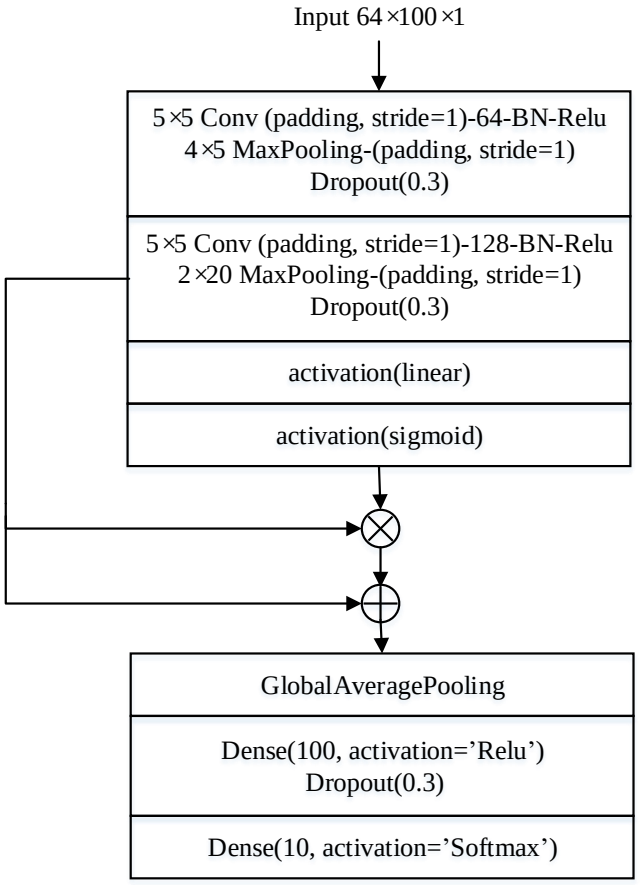

Figure 3: Proposed Neural Network Architecture.

Table 1: Results

Network	Division	Accuracy(%)
Baseline	Non	62.50
CNN	Overlap	70.20
CNN+attention	Overlap	70.70

4. CONCLUSION

In this report, we first divide the spectrogram into smaller segments to be used as the input of the network model. This method can achieve great classification performance compared with the baseline system. In order to further improve the performance, we propose an attention-based network architecture along frequency dimension. We perform the experiments on the dataset of DCASE 2019 Task1 Subtask A. Experimental results illustrate that the best classification performance can be acquired of the network model with overlap division method combining with attention mechanism. In future, our aim is to pursue a better acoustic feature representing the raw audio signal, which can help to achieve better classification results.

5. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 61672246, No. 61272068. And we would like to acknowledge NVidia Corporation for the donation of Titan XP GPUs for this research.

Table 2: Confusion matrix of the best model

	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
airport	306	74	18	18	0	3	2	0	0	0
mall	49	302	33	44	12	0	0	0	0	1
met stn	16	44	265	19	4	9	11	14	53	0
st ped	15	6	16	313	73	4	0	0	1	1
pub sq	9	4	5	73	224	28	10	3	3	28
st trf	3	0	2	10	44	336	0	1	0	6
tram	0	2	10	9	0	0	248	102	65	0
bus	0	0	10	0	0	1	40	336	26	2
metro	6	4	28	2	1	1	28	52	311	0
park	0	0	12	5	19	30	2	0	0	318

6. REFERENCES

[1] M. Dorfer, B. Lehner, H. Eghbal-zadeh, H. Christop, P. Fabian, and W. Gerhard, "Acoustic scene classification with fully convolutional neural networks and i-vectors," *IEEE AASP Challenge on Detection and Classification of Acoustic Scen and Events (DCASE)*, 2018.

[2] L. Vuegen, B. Broeck, P. Karsmakers, J. F. Gemmeke, B. Vanrumste, and H. Hamme, "An mfcc-gmm approach for event detection and classification," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–3.

[3] H. Zeinali, L. Burget, and J. Cernocky, "Convolutional neural networks and x-vector embedding for dcase2018 acoustic scene classification challenge," *arXiv preprint arXiv:1810.04273*, 2018.

[4] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," *IEEE AASP Challenge on DCASE 2018 technical reports*, 2018.

[5] J. Wang and S. Li, "Self-attention mechanism based system for dcase2018 challenge task1 and task4," *IEEE AASP Challenge on DCASE 2018 technical reports*, 2018.

[6] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and svm for acoustic scene classification," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[7] K. Qian, Z. Ren, V. Pandit, Z. Yang, Z. Zhang, and B. Schuller, "Wavelets revisited for the classification of acoustic scenes," in *Proc. DCASE Workshop, Munich, Germany*, 2017, pp. 108–112.

[8] A. Vafeiadis, D. Kalatzis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Acoustic scene classification: From a hybrid classifier to deep learning," 2017.

[9] S. S. R. Phaye, E. Benetos, and Y. Wang, "Subspectralnet–using sub-spectrogram based convolutional neural networks for acoustic scene classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 825–829.

[10] Z. Ren, Q. Kong, J. Han, M. Plumbley, and B. W. Schuller, "Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes," in *2019 Proceedings IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP 2019). IEEE, 2019.

[11] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," *arXiv preprint arXiv:1807.09840*, 2018.