ACOUSTIC SCENE CLASSIFICATION USING NETWORK-IN-NETWORK BASED CONVOLUTIONAL NEURAL NETWORK

Andri Santoso, Chien-Yao Wang, Jia-Ching Wang

National Central University, Taiwan

ABSTRACT

In this paper, we present our entry to the challenge of detection and classification of acoustic scenes and events (DCASE). This paper describes the result of our proposed system for automatic audio scene classification task. Our approach is based on the deep learning method that is adopted from computer vision research field. The convolutional neural network is adopted to solve the problem of audio based scene classification, specifically the architecture of network-in-network is utilized to build the classifier. For the feature extraction part, mel frequency spectral coefficients (MFCC) is used as the input vector for the classifier. Differ from the original architecture of network-in-network, in this work we perform 1-D convolution operation instead of performing 2-D convolution. The classifier is trained using every frames from MFCC feature set, and the results for every frames are then thresholded and voted to choose the final scene label of audio data. The proposed work in this paper shows a better performance of the provided baseline system of DCASE challenge for both development and evaluation dataset.

Index Terms— acoustic scene classification, sound event recognition

1. INTRODUCTION

In the era of emerging mobile devices, an ability of a device to sense its environment is very beneficial for the users to have a better use of their smart-phone. Enabling the device to sense its environment is related to the research works of scene classification. Some research works address the problem of scene classification using visual-based approaches [1, 2]. Even though visual-based approach has a promising performance to address the problem of scene classification, its performance is compromised when the visual information is fully or partially blocked, thus it reduces the flexibility of the approach. The analysis of sound is promising alternative to address this issue. As long as the device can ”listen” to the sound, a practical acoustic scene classification (ASC) system should be able to analyze the audio data and detect the environment where it lives. The analysis of audio data to detect scene information is related to the works of machine listening, as it enables the machine to understand the audio data that it listened.

In this work, an ASC system is proposed to solve the problem of automatic audio scene classification using the approach that is adopted from the research of computer vision. In recent years, the audio processing research field has seen an increasing interest of adopting the methods from computer vision field to perform audio analysis.

Furthermore, the recent advances in machine learning has seen a growing interest in the use of deep learning based approaches. One of the architecture that received a lot of interest is convolutional neural network. Convolutional neural network (CNN) [4] scans the input using patch-based approach and learns the visual representation of input image in every layers of the network. Most of the work using CNN is in the computer vision field, as it has shown a tremendous success in the practical applications in that field [4, 5]. To enhance the discrimination ability of the CNN for local patches within receptive field, Lin et al [6] proposed an architecture of network-in-network that introduces a micro neural network inside convolutional layers to abstract the data within receptive field.

Motivated by above mentioned works, we designed a classification system, based on the network-in-network architecture to address the problem of audio based scene classification. Differ from the original work of network-in-network, we perform 1-D convolution operation on the input data instead of performing 2-D convolution operation. We performed 1-D convolution in every extracted frames from MFCC features set. The filter then scans 1-D vector and result the output of convolution operation as the feature maps. The architecture of the neural network in this work is illustrated in Fig. 1.

2. PROPOSED METHOD

In this work, the automatic audio scene classification system is proposed. The system performs training process of the classifier based on the architecture of network-in-network [6], and use the extracted MFCC feature as the input. This section will first discuss about the underlying components of our approach and then detail the proposed classification system.

2.1. Convolutional Neural Network

The convolutional neural network (CNN) [4] has presented a promising results in the computer vision research field. In its basic form, CNN consists of two layers, which are convolutional and pooling layers. The convolution operation that is used in CNN is listed in Eq. 1.

$$S(i,j) = (I * K)(i,j) = \sum \sum I(m,n)K(i-m, j-n) \quad (1)$$

where I is the image input, and K is the kernel that is used to perform convolution operation.

The convolution and pooling layers are stacked together to scan image input and then generate the feature maps. Then a downsampling operation is performed to the feature maps using pooling operation. The last layer of CNN is connected to a multi-layer perceptron (MLP) neural network that acts as a classifier.

The advance of CNN model has presented a number of new architectures to construct a classifier using deep learning based approach. One of the proposed architecture is network-in-network (NIN) architecture [6].

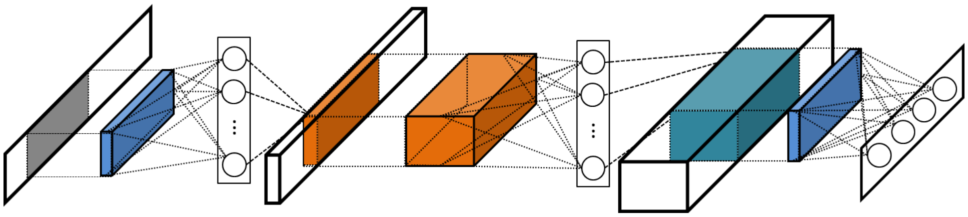

Figure 1: The neural network architecture of proposed audio scene classification system. The input is 1-D vector and the convolution operation in every layer of convolutional network is 1-D convolution operation. The output of the network is the scene labels in the dataset.

2.2. Network-in-network architecture

The network-in-network (NIN) [6] architecture is proposed to improve the abstraction ability of local model in the convolution layer of CNN. NIN replaces the model that perform the abstraction on data patch in the CNN with a more potent nonlinear approximator. In the NIN architecture, the abstraction model is replaced by an MLP network.

Furthermore, the NIN architecture replaces the classification approach of traditional CNN. In the CNN architecture, the feature maps are connected to the traditional MLP network that acts as the classifier. The NIN architecture replaces this classification approach using global average pooling. The NIN architecture uses directly the feature maps in the last convolution layer to build the classifier. This architecture takes the average of the feature maps, and the resulting vector is fed directly into the softmax layer.

2.3. Proposed classification system

This work adopted the NIN architecture to train the classifier. Differ from the original implementation of NIN in [6], the proposed system performs 1-D convolution operation, instead of 2-D convolution. The proposed system perform convolution operation on the vector that is obtained from every frames in the feature set.

As is shown in the Fig. 1, the input for the neural network is a vector that is obtained from every frames in the extracted MFCC feature set. Then, filters with depth n are defined to perform the convolution operation with the patches from input vector.

The neural network architecture in this work is consisted of five layers with the filter that has the depth of 128, 128, 64, 64, and 64 for the convolutional layer 1, 2, 3, 4, and 5, respectively. Following the work of NIN model, we connected the last feature maps into the softmax layer to perform the training and classification task.

As we use frame-based approach to give the scene labels for each audio file, we applied a thresholding approach to the output from the neural network. Then, using the thresholded result, we perform the voting mechanism to choose the scene labels for the particular audio file.

3. EXPERIMENTAL SETUP

The TUT acoustic scene 2016 dataset [7] is used to verify the performance of the proposed system. The dataset is consisted of audio data that is recorded in various environments.

4. EXPERIMENTAL RESULTS

The performance evaluation of the proposed system is done by comparing the result of provided DCASE baseline system [7], with the result of our proposed system. The approach in the baseline system uses the Gaussian mixture model as the classifier and MFCC as the feature.

Table 1 shows the performance comparison of our method with the one provided from the DCASE challenge using the development dataset from TUT acoustic scene 2016 dataset. Besides the accuracy for fold 4, the performance of our proposed system exceeds all the performances of baseline system.

For the average accuracy, our proposed system outperformed the baseline system of DCASE challenge. Our proposed system obtained the average accuracy of 78.83%, while the baseline system obtained 72.57% average accuracy.

Table 1: Performance of the systems for all folds in development dataset.

Approach	Fold1	Fold2	Fold3	Fold4
Baseline system [7]	67.2%	68.9%	72.3%	**81.9%**
Proposed system	**82.41%**	**79.66%**	**75.84%**	77.4%

In the Table 2, all performances of proposed and baseline system using the evaluation dataset for DCASE challenge are provided. The proposed system outperformed the baseline by 3.6%. The results showed that the performance of our ASC system exceeds the baseline system in both development and evaluation dataset.

Table 2: Performance of the systems in evaluation dataset.

Class	Baseline	Proposed System
Beach	84.60%	84.60%
Bus	88.50%	84.60%
Cafe /Restaurant	69.20%	61.50%
Car	96.20%	96.20%
City center	80.80%	84.60%
Forestpath	65.40%	100%
Grocerystore	88.50%	80.80%
Home	92.30%	100%
Library	26.90%	42.30%
Metro Station	100%	92.30%
Office	96.20%	100%
Park	53.80%	80.80%
Residential area	88.50%	65.40%
Train	30.80%	42.30%
Tram	96.20%	96.20%
Average	77.20%	**80.80%**

5. CONCLUSIONS

This paper summarizes our entry for the DCASE challenge. In this work, a deep learning based system is proposed to perform automatic audio based scene classification. The system is designed based on the architecture that is adopted from the work in the computer vision field, specifically the network-in-network architecture is utilized to build the classifier in this work. The experimental results show that our proposed system outperformed the baseline system that is provided for the challenge. The proposed system showed the superiority over the baseline for both development and evaluation dataset for DCASE challenge.

6. REFERENCES

[1] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, June 2005, pp. 524–531 vol. 2.

[2] A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, April 2008.

[3] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[4] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.

[5] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceesdings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1701–1708.

[6] M. Lin, Q. Chen, and S. Yan, "Network in network," *CoRR*, vol. abs/1312.4400, 2013. [Online]. Available: http://arxiv.org/abs/1312.4400

[7] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *Proceedings of 24th European Signal Processing Conference 2016 (EUSIPCO 2016)*, Budapest, Hungary, 2016.