A MULTI-SCALE DEEP CONVOLUTIONAL NEURAL NETWORK FOR ACOUSTIC SCENE CLASSIFICATION

Jianhao Wang

Tsinghua University
Institute for Interdisciplinary Information Sciences
1040594377@qq.com

Tao'an Huang

Tsinghua University
Institute for Interdisciplinary Information Sciences
hta15@mails.tsinghua.edu.cn

ABSTRACT

Deep neural networks have shown great classification performances in numbers of applications. We applied a multi-scale deep convolutional neural network to acoustic scene classification (ASC) which has been submitted to Task 1 of the DCASE-2017 challenge. In this report, we show our model for classifying short sequences of audio, represented by their Mel-Frequency Cepstral Coefficients and Constant-Q value. The system is evaluated on the public dataset provided by the organizers. The best accuracy we obtained on a 4-fold cross-validation setup is 84.4%.

***Index Terms*—** acoustic scene classification, convolutional neural network, multi-scale

1. INTRODUCTION

ASC is a task to understand the context of an audio and to identify a particular acoustic environment. In the last few years, many models have been proposed to deal with this task in artificial system. The first work introducing a CNN-based classifier for ASC was submitted to DCASE-2016[2]. In this report, we propose the use of multi-scale deep convolutional neural network trained to classify audio sequences.

In section 2, a detailed description of our proposed method for this task is given which includes the extraction of features and the architecture of the multi-scale deep neural network model. In section 3, we describes the experiment setup and the methods we uses for training and evaluation.

2. PROPOSED METHOD

In this section, we will introduce our extraction of features, the architecture of the deep neural network and the method of training.

2.1. Feature Extraction and Preprocessing

The feature we choose for our system is Mel-Frequency Cepstral Coefficients (MFCC), a frame-level feature, and Constant-Q Chroma Gram.

To calculate MFCC, we apply short-time Fouries transform twice over FFT window size of 2048 and 8192 respectively. Then we square the absolute value of those coefficients, compute mel-scaled spectrograms and convert them into decibel units. Finally, we apply a MFCC computations with 64 coefficients to return.

As for Constant-Q transform (CQT), we simply apply a chroma CQT with 64 chroma bins to produce.

2.2. Proposed Architecture

The proposed model is represented in Figure 1.

The model starts with 3 branches. The first stack of layers of each branch performs two convolution with 64 kernels characterized by 5×5 receptive field using rectifier function as the activation function, and then followed by a Batch Normalization (BN) layer

The second stack of layers of each branch subsamples the obtained feature map to the same scale and concatenate features from all branches.

The third stack is the same as the first one, except there are a Max-Pooling along time axis and a Dropout with probability 0.3 instead of a BN layer.

Finally, we flatten the feature map and connect it to a dense layer with 128 units. The last is a softmax layer composed of 15 fully-connected neurons for the purpose of 15-class classification.

3. EXPERIMENT

3.1. Dataset

We executed an evaluation of proposed model with the dataset provided by DCASE 2017 challenge. They provided 4 folds of dataset for evalution. All audio are recorded with the sampling rate of 44.1kHz. We split each given 10-second recording into 13 non-overlapping segments.

3.2. Training

We choose categorical cross entropy between target label as our loss function of all of the 3 branches. The total loss is the summation of the 3 losses with uniform weight. We choose Adam as our optimizer for faster fitting and the batch size is set as 64.

The predicted class of an audio is given by the majority of the results of 13 audio segments split from the orginal audio.

Table 1: Comparing the perfomance of our method with the baseline provided by the DCASE organizers. Row Base: the performance of the baseline. Row Ours: the performance of our method.

	Mean accuracy (%)				
	Fold 1	Fold 2	Fold 3	Fold 4	Avg.
Base	76.2	73.4	78.1	72.6	75.1
Ours	82.4	83.9	82.3	89.1	84.4

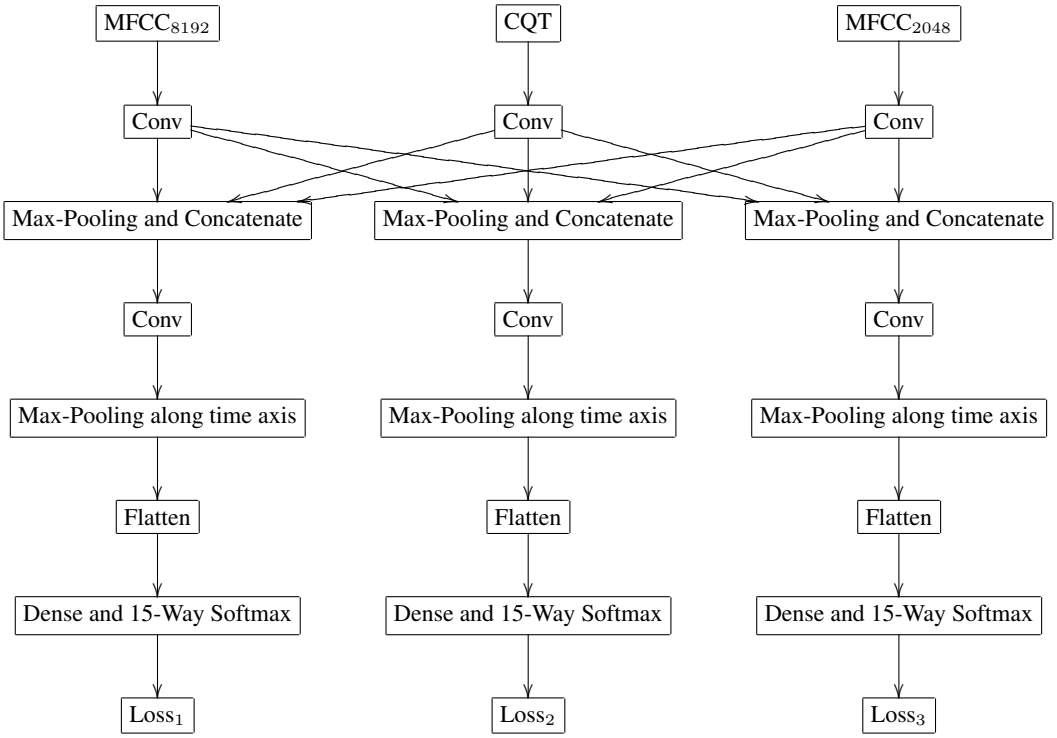

Figure 1: Block diagram of the proposed architecture of the multi-scale neural network.

3.3. Evaluation

For evaluation, we trained 4 models out of 4 folds of dataset. For each audio, we get 13 answers from each model and hence result in a total of 52 answers. We take the majority of those 52 answers as the predicted class for this audio.

4. CONCLUSION

Our work proposes a way of approaching ACS with multi-scale deep convolutional neural networks. We achieve an accuracy of 84.4% on the DCASE-2017 development dataset, which constitute a 9.3% relative improvement with respect to the baseline.

5. REFERENCES

[1] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, submitted.

[2] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks," DCASE2016 Challenge, Tech. Rep., September 2016.