ACOUSTIC SCENE CLASSIFICATION USING ENSEMBLE OF CONVNETS

Technical Report

An Dang, Toan Vu, Jia-Ching Wang

Department of Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan

ABSTRACT

This technical report presents our system for the acoustic scene classification problem in the task 1A of the DCASE2018 challenge whose goal is to classify audio recordings into predefined types of environments. The overall system is an ensemble of ConvNet models working on different audio features separately. Audio signals are processed in both mono channel and two channels before we extract mel-spectrogram and gammatone-based spectrogram features as inputs to models. All models are implemented by almost the same ConvNet structure. Experimental results illustrate that the ensemble system can achieve superior accuracy to the baseline by a large margin of 17% on the test data.

Index Terms— Acoustic scene classification, ConvNet, Ensemble

1. ACOUSTIC SCENE CLASSIFICATION SYSTEM

The acoustic scene classification (ASC) task [1] in the DCASE2018 challenge is to recognize ten environmental scenes including - *airport*, *bus*, *metro*, *metro station*, *park*, *public square*, *shopping mall*, *street pedestrian*, *street traffic*, and *tram*. Recordings are collected from different locations by several devices. They can be very noisy, or sound very similar between scenes, which make the task difficult. In this challenge, we mainly focus on solving the ASC problem with recordings from the same device.
Similar to previous works [2, 3, 4], our system is an ensemble of models that are trained separately on different audio features. The overall ASC system is illustrated in Figure 1.

1.1. Audio processing

Following the work in [2], audio signals are processed in both mono and stereo. Log mel-spectrogram and gammatone based spectrogram are employed as audio features with the following parameters - sampling rate of 48 KHz, window size of 2048, hop size of 1024, and number of mel filter-banks and number of gammatone filter-banks is 128. 10-seconds audio files are divided to smaller segments with length of 2 seconds to be used as inputs of models.
For mono signals, on one branch, mel-spectrogram and gammatone-spectrogram are extracted before we do background subtraction on these features by using a median filter with size of 21, 11 on time axis and frequency axis, respectively. On the other branch, we apply harmonic-percussive source separation on the mono signals to extract harmonic components and percussive components before extracting spectrograms from them.
For stereo signals, we simply extract spectrograms directly from left and right channels on one branch, while on the other branch, the two channels are combines by addition and subtraction before extracting features.

Table 1: Acoustic scene classification results on the test set

Label	Accuracy (%)	
	Baseline	**Ours**
Airport	72.9	81.5
Bus	62.9	69.8
Metro	51.2	77.8
Metro station	55.4	82.2
Park	79.1	84.7
Public square	40.4	55.1
Shopping mall	49.6	75.6
Street, pedestrian	50.0	69.6
Street, traffic	80.5	91.1
Tram	55.1	79.7
Average	**59.7**	**76.7**

1.2. ConvNet models

We implement the same ConvNet structure (conv_net) for all input features as can be seen in Figure 1&2. In case of stereo signals, outputs of two input features after the convnet are concatenated before going through fully connected layers. Figure 3 displays the structure the convnet. Construction of the convnet is based on residual network [5].
We train models on their corresponding input features independently, which results in eight models. After training process, predictions from all models are computed by averaging to make a final prediction.

2. EXPERIMENTAL RESULTS

The TUT Urban Acoustic Scenes 2018 dataset in the task 1A contains recordings collected from various locations by using the same device. There are totally 10 classes - *airport*, *bus*, *metro*, *metro station*, *park*, *public square*, *shopping mall*, *street pedestrian*, *street traffic*, and *tram*. From our observation, the task is particularly hard even for people with blind-hearing test because the recordings are noisy, and it is very easy to be confused between scenes. The training data has 6122 files, the test data has 2518 files. In order to improve accuracy, we randomly divide the training data into two parts that combine with the test data to get a 3-folds

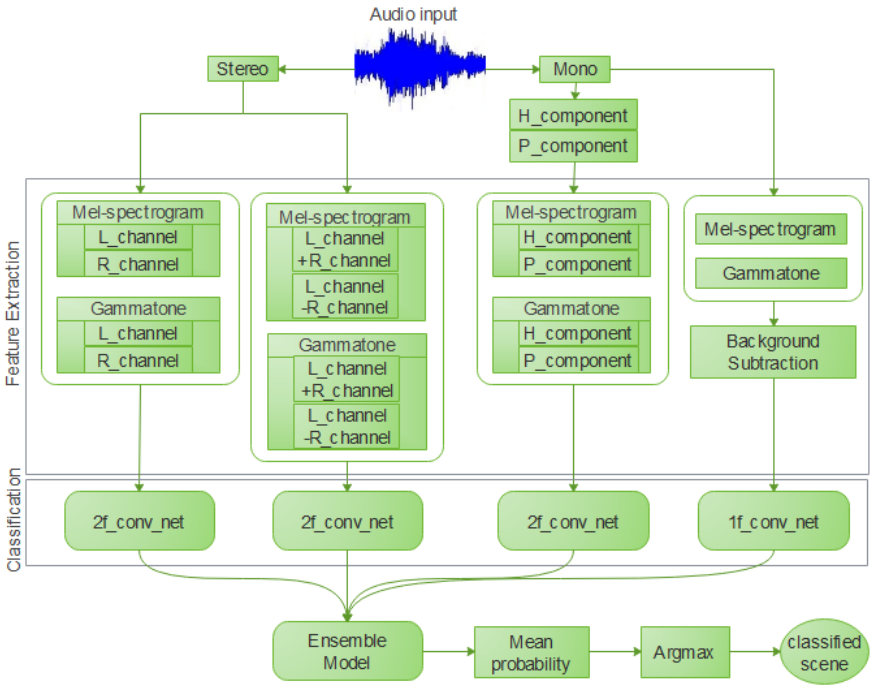

Figure 1: The acoustic scene classification system.

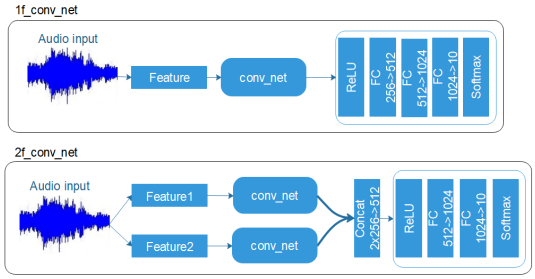

Figure 2: Network structures for different input features.

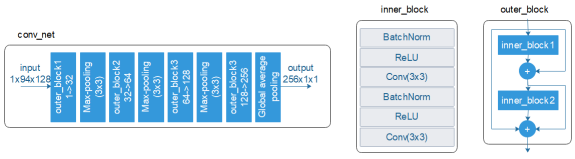

Figure 3: ConvNet architecture.

cross-validation setting.

Table 1 show our classification results on the test set with the given evaluation setting. We achieve an average accuracy of 76.7% which is superior to that of the baseline system by a large margin of 17%. The *street traffic* scene is the most recognizable, while the *public square* scene is the hardest one to classify. In our results, many *public square* and *shopping mall* samples are misclassified as *street pedestrian* and *airport*, respectively.

3. CONCLUSION

In this report, we introduce our experimental results for the task 1A of acoustic scene classification in the DCASE 2018 challenge. By ensembling convnet models on various audio features, our system can achieve significant accuracy.

4. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," 2018, submitted to DCASE2018 Workshop. [Online]. Available: https://arxiv.org/abs/1807.09840

[2] Y. Han and J. Park, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," DCASE2017 Challenge, Tech. Rep., September 2017.

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

[4] Z. Weiping, Y. Jiantao, X. Xiaotao, L. Xiangtao, and P. Shaohu, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion," DCASE2017 Challenge, Tech. Rep., September 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385