ACOUSTIC SCENE CLASSIFICATION USING ENSEMBLES OF CONVOLUTIONAL NEURAL NETWORKS AND SPECTROGRAM DECOMPOSITIONS

Technical Report

Shengwang Jiang, Chuang Shi, Huiyong Li

University of Electronic Science and Technology of China, Chengdu, China
swjiang@std.uestc.edu.cn, shichuang@uestc.edu.cn, hyl@uestc.edu.cn

ABSTRACT

This technical report proposes ensembles of convolutional neural networks (CNNs) for the task 1 / subtask B of the DACSE 2019 challenge, with emphasis on using different spectrogram decompositions. The harmonic percussive source separation (HPSS), nearest neighbor filter (NNF), and vocal separation are applied to the monaural samples. Head-related transfer function (HRTF) is also proposed to transform monaural samples to binaural ones with augmented spatial information. Finally, 16 neural networks are trained and put together. The classification accuracy of the proposed system achieves 0.70166 on the public leaderboard.

Index Terms— DACSE 2019, acoustic scene classification, convolutional neural network, spectrogram decomposition, HRTF

1. INTRODUCTION

Acoustic scene is defined as the environment in which an audio clip has been recorded. Acoustic scene classification aims to recognize the acoustic scene from an audio clip [1]. Human beings are capable to do so based on auditory perceptions. However, the classification accuracy is not satisfactory. With the rapid development of machine learning techniques, computer audition has gradually gained a similar capability of recognizing the environments where they are being used by analyzing the patterns in the sound [12]. The vocal separation divides a mixture into the non-repeating foreground and repeating background sounds [13]. Besides spectrogram decompositions, the HRTF is another useful tool, which defines the transfer function from a sound source in a spatial position to ears of a listener [14]. Multi-channel output can be created by HTRFs with a monaural input.

To augment the features, spectrogram decompositions are introduced. This technical report selects the HPSS, NNF, vocal separation. The HPSS decomposes a monaural audio into two channels: one contains the harmonic sounds and other contains the percussive sounds [11]. The NNF emphasizes and smooths the spectrum of the sound [12]. The vocal separation divides a mixture into the non-repeating foreground and repeating background sounds [13].

2018, the task of acoustic scene classification has been divided into 3 subtasks. Among them, the subtask A works on the dataset collected with a single device. In 2019, the classification accuracy of the subtask A can achieve higher than 0.86 on the public leaderboard. Closer to the practical situation, the subtask B provides the dataset from multiple devices. Particularly, some devices are only used to record the evaluation dataset. Currently, the classification accuracy of the subtask B is significantly lower than that of the subtask A [4].

When a machine learning application is developed, a large number of structured data often benefits the classification accuracy [5]. The dataset of the DCASE 2019 challenge, although getting larger than the previous year’s, is yet to be sufficient. Therefore, the network architecture and feature extraction still require careful selection.

Through intensive competition, the CNN has won the most widely used architecture in the task of acoustic scene classification during the DCASE 2018 challenge [6]. The convolution incurred in the CNN fuses the cross-frame and cross-frequency-bin information, which results in a high-level retrieval of features. As compared to the deep neural network (DNN), the scale of CNN is more concise and thus requires less data for training [7]. The recursive neural network (RNN) is advantageous in dealing with temporal sequence. However, the long short-term memory (LSTM) has not achieved higher accuracy than the baseline in the subtask A of acoustic scene classification in the DCASE 2018 challenge [8].

The most common features used in acoustic scene classification are the spectrogram and its variants, including the short-time Fourier transform (STFT), log mel spectrogram, mel frequency cepstral coefficients (MFCC), constant-Q transform (CQT). Among them, the log mel spectrogram is found to provide the best performance in the DCASE 2018 challenge [10].

This work is jointly supported by the National Science Foundation of China (Grant No. 61671137 and Grant No. 61701090) and Sichuan Science and Technology Program (Project No. 2018JY0218 and 2019JY0182).
any of the individual model. Ensembles of CNNs have been attempted in the previous DCASE challenges by many participants [11]. The variety in the spectrogram decompositions can improve the generalization of the ensembles of CNNs.

2. ARCHITECTURE

2.1 Network Architecture

The CNN we use in this technical report was originally proposed by Han et al and inspired by the VGGNet [15]. In total, 4 convolution blocks are used. Each of the convolution block consists of 2 convolution layers. Batch normalization (BN) is adopted instead of dropout [16]. In order to accelerate the convergence of model training, we perform a 3×3 max pooling at the end of each convolution block (see Figure 2) and find this practice is beneficial to the classification accuracy. After the convolution blocks, the global average pooling is preferable in terms of the classification accuracy. However, the global max pooling is faster than the global average pooling. Therefore, we have used both of them in the ensembles.

2.2 Spectrogram Decomposition

2.2.1 Harmonic percussive source separation (HPSS)

The HPSS treats an audio input as a combination of the harmonic and percussive components. The HPSS was initially developed to separate the drums from a mixture by using the median filter. The Librosa package has provided a simple function to carry out the HPSS.

2.2.2 Nearest neighbor filter (NNF)

The NNF removes the outliers. Therefore, it smooths the features to focus more on the overall picture instead of the details. Previous works have validated the positive effect of using the NNF in the task of acoustic scene classification [4]. The Librosa package provides two options to do the NNF. One option is with non-local means method by setting ‘aggregate’ to ‘np.average’. The other option sets ‘aggregate’ to ‘np.median’. The latter option is chosen in this technical report.

2.2.3 Vocal separation

The vocal separation is a technique for separating the vocal sound from the accompanying instrumentation. In the task of acoustic scene classification, the vocal separation is used to separate the sporadic foreground signal from the background signal with certain patterns. The background information is more likely to reflect the acoustic scene. Similarly, we use the Librosa package to implement the vocal separation with default settings.

2.3 HRTF

The HRTF is the transfer function that describes the process of an ear receiving sound from a point in space. Since human beings have two ears, the HRTF appears in a pair. We introduce the HRTF to preprocess a monaural sample and result in a binaural sample. In the CIPIC database, HRTFs are provided at a number
of azimuth and elevation angles [17]. The front and back positions are chosen in this technical report, with the inspiration from a psychoacoustic effect called the front-back confusion.

### 2.4 Features

The samples in the DCASE2019 task 1 / subtask B dataset are monaural and have a common sampling rate of 44.1 kHz. Each sample or each channel of a preprocessed sample can generate one spectrogram by using the 2048-point hamming window with the hop size of 1024 samples. The log mel spectrogram is implemented by applying the log mel filter bank on the spectrogram. There are 128 log mel filters in the filter bank and together they cover a frequency range from 0 to 22.05 kHz. The log mel spectrograms are standardized by subtracting the mean and dividing the standard deviation.

Therefore, the output of the NNF is combined with the input to generate the \((431, 128, 2)\) feature ‘Mnf_2ch’. The output of the HPSS results in a two-channel feature ‘Mhp_2ch’ with the size of \((431, 128, 2)\) and together with the input forms a three-channel feature ‘Mhp_3ch’ with the size of \((431, 128, 3)\). Similarly, the vocal separation also provides the \((862, 128, 3)\) feature ‘Mbf_3ch’ and the \((862, 128, 2)\) feature ‘Mbf_2ch’ with and without the input, respectively. Note that the hop size setting in the vocal separation is different from that of the HPSS. This results in a change of the feature size. Moreover, we obtain the ‘Hrtf_2ch’ and ‘Hrtf_4ch’ features when only the front position is simulated and both the front and back positions are simulated by the HRTFs, respectively.

### 2.5 Network Ensemble

The CNN models are trained individually and then put together to make the final decision, whereby the decision making strategy is also learnt from data. The ensemble of CNNs achieves higher classification accuracies and better generalization. The common methods of ensemble include voting, averaging, weighted averaging, and stacking. We choose the averaging and stacking for comparison. The averaging method averages the output probabilities of different models, as shown in Figure 3 when all the weights are equal. The stacking is often a more effective method. We choose the CNNs to be the base learners and the random forest to be the meta-learner, as illustrated in Figure 4.

### 3. EXPERIMENTS

#### 3.1 Dataset

The dataset of the task 1 / subtask B of the DCASE 2019 challenge consists of 10 acoustics scenes. They are the airport, shopping mall, metro station, street pedestrian, public square, street traffic, tram, bus, metro, and park. The development dataset consists of data recorded with 3 devices. They are referred to as the devices A, B, and C. The device A is also used to record the dataset of the subtask A. It contributes the longest hours of data in the whole dataset. The evaluation dataset contains audio samples recorded by 4 devices, including an additional device D that is not used to record the development dataset.

#### 3.2 Results and Submissions

The stochastic gradient descent (SGD) using Nesterov momentum is adopted for the model training. The learning rate, decay, and momentum are set to 0.01, 0.0001, and 0.9, respectively. Each model is trained for 15000 iterations and takes 1.5-2.5 hours on one NVIDIA GTX 980Ti card. In the random forest, the number of decision trees are set to 5000.

Table 1 lists the models that we submit. All the four submissions achieve higher classification accuracies in the public leaderboard dataset than in the development dataset.
Table 1: Results of development and public leaderboard dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Development</th>
<th>Public Leaderboard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.414</td>
<td>0.43833</td>
</tr>
<tr>
<td>Averaging_8</td>
<td>0.640</td>
<td>0.65333</td>
</tr>
<tr>
<td>Averaging_16</td>
<td>0.632</td>
<td>0.68666</td>
</tr>
<tr>
<td>Randomforest_8</td>
<td>0.642</td>
<td>0.69166</td>
</tr>
<tr>
<td>Randomforest_16</td>
<td>0.622</td>
<td>0.70166</td>
</tr>
</tbody>
</table>

- Averaging_8 is the averaging ensemble of 8 models that perform the global average pooling.
- Averaging_16 is the averaging ensemble of 16 models that perform both the global average pooling and global max pooling.
- Randomforest_8 is the stacking ensemble of 8 models that perform the global average pooling.
- Randomforest_16 is the stacking ensemble of 16 models that perform both the global average pooling and global max pooling.

4. CONCLUSIONS

In this paper, we ensemble 16 models to improve the accuracy of the acoustic scene classification. Three spectrogram decompositions and the HRTF are proposed to augment the acoustic features. Both the averaging and stacking are considered for the ensemble. The results show that these methods can improve the classification accuracy as compared to the baseline.

5. REFERENCES


