ACOUSTIC SCENE CLASSIFICATION USING CNN ENSEMBLES AND PRIMARY AMBIENT EXTRACTION

Technical Report

Haocong Yang[1], *Chuang Shi*[2], *Huiyong Li*[3]

University of Electronic Science and Technology of China, Chengdu, China
[1] yanghaocong@std.uestc.edu.cn
[2] shichuang@uestc.edu.cn
[3] hyli@uestc.edu.cn

ABSTRACT

This report describes our submission for Task 1a (acoustic scene classification) of the DCASE 2019 challenge. The results of the DCASE 2018 challenge demonstrate that the convolution neural networks (CNNs) and their ensembles can achieve excellent classification accuracies. Inspired by the previous works, our method continues to work on the ensembles of CNNs, whereas the primary ambient extraction is newly introduced to decompose a binaural audio sample into four channels by using the spatial information. The feature extraction is still carried out with mel spectrograms. 6 CNN models are trained by using the 4-fold cross validation. Ensemble is applied to further improve the performance. Finally, our method has achieved classification accuracies of 0.84 on the public leaderboard.

***Index Terms*—** DCASE 2019, acoustic scene classification, convolutional neural network, primary ambient extraction.

1. INTRODUCTION

Sound carries a lot of information and plays an important role in our everyday life [1]. We receive various kinds of sound and use them to judge where we are (metro, street, etc.) and what is happening (sirens, dog barking, etc.). Those two types of answers are therefore called the acoustic scene and acoustic event, respectively. With the rapid development of artificial intelligence, such judgements can also be made by computers, whose accuracies even surpass those of human beings. The computer audition and machine listening become a popular and promising research frontier [2].

The DCASE challenge aiming to extend the start-of-the-art of acoustic scene and event analysis methods has been organized for more than 6 years [3]. As a regular task of the DCASE challenge, the competition on acoustic scene classification has been held for four times. In 2018, the task of acoustic scene classification (task 1) was divided into 3 subtasks [4]. Each of them concentrated on a specified perspective derived from different real-world requirements. Subtask A continued to focus on the classic scope of acoustic scene classification. Subtask B emphasized the difficulty in acoustic scene classification when different devices were adopted to record the development and evaluation datasets. Subtask C expanded the number of acoustic scenes in the way that the evaluation dataset included new classes not encountered in the development dataset.

A successful machine learning application requires a high-quality dataset [5]. The DCASE challenge improves the dataset year by year. In 2016, the audio samples in the development and evaluation datasets lasted for 30 s. There were 78 audio samples in each acoustic scene. In 2017, the audio samples were divided into audio segments lasting for 10 s and each acoustic scene corresponded to 312 segments consequently. The evaluation dataset was recorded separately from the development dataset. In last year, high-quality binaural audio samples were recorded in 6 cities in Europe. However, the number of acoustic scenes was reduced from 15 to 10, in order to have more audio samples representing every acoustic scene. In this year, the same recording devices are used in another 6 cities in Europe in order to further improve the variety.

The results of the previous DCASE challenges suggest that CNNs are the most popular classifiers for acoustic scene classification [6][7]. With the special structure of local weight sharing, the CNN requires less data than the deep neural network (DNN) [8]. Models with recurrent neural network (RNN) or long short-term memory (LSTM) are more suited to temporal sequences. They are yet to obtain satisfactory accuracies in the DCASE challenge till now [9]. On the other hand, a number of features, such as the constant-Q transform (CQT), mel frequency cepstral coefficients (MFCC), and etc., have been attempted [10][11]. Among them, the mel spectrogram remains to be the most preferable [12][13]. Ensemble is a widely applied approach among the top teams [14][15]. Therefore, data preprocessing has to be paid attention to, in order for more effective features to be generated for ensemble.

In this report, we introduce the primary ambient extraction in data preprocessing. Every binaural audio sample is separated into four channels, which are two primary channels and two ambient channels. By doing so, spatial information is preserved, such that the phase variation is contained in the mel spectrogram to a certain extend. The cropped and raw mel spectrograms are used to train CNN models with different structures. Finally, we ensemble the trained models to refine the classification accuracy by using a random forest.

Thanks the National Science Foundation of China (Grant No. 61701090 and Grant No. 61671137) and Sichuan Science and Technology Program (Project No. 2019YJ0182 and 2018JY0218) for supporting this work.

Input 431×128×1 or 431×128×2 or 431×128×4
7×7 Conv2D (pad=1, stride=1)-32-BN-ReLU 7×7 Conv2D (pad=1, stride=1)-32-BN-ReLU 2×2 MaxPooling2D
3×3 Conv2D (pad=1, stride=1)-64-BN-ReLU 3×3 Conv2D (pad=1, stride=1)-64-BN-ReLU 2×2 MaxPooling2D
3×3 Conv2D (pad=1, stride=1)-128-BN-ReLU 3×3 Conv2D (pad=1, stride=1)-128-BN-ReLU 5×5 MaxPooling2D
3×3 Conv2D (pad=1, stride=1)-256-BN-ReLU 3×3 Conv2D (pad=1, stride=1)-256-BN-ReLU GlobalAveragePooling2D
Dense (512, activation='relu')
Dense (10, activation='softmax')

Figure 1: CNN structure for the raw mel spectrogram features

Input 129×128×1 or 129×128×2 or 129×128×4
3×3 Conv2D (pad=1, stride=1)-32-BN-ReLU 3×3 Conv2D (pad=1, stride=1)-32-BN-ReLU 2×2 MaxPooling2D
3×3 Conv2D (pad=1, stride=1)-64-BN-ReLU 3×3 Conv2D (pad=1, stride=1)-64-BN-ReLU 2×2 MaxPooling2D
3×3 Conv2D (pad=1, stride=1)-128-BN-ReLU 3×3 Conv2D (pad=1, stride=1)-128-BN-ReLU 5×5 MaxPooling2D
3×3 Conv2D (pad=1, stride=1)-256-BN-ReLU 3×3 Conv2D (pad=1, stride=1)-256-BN-ReLU GlobalAveragePooling2D
Dense (512, activation='relu')
Dense (10, activation='softmax')

Figure 2: CNN structure for the cropped mel spectrogram features

2. DATA PREPROCESSING

This section describes the signal processing and data augmentation methods that we implement to transform an audio sample into acoustic features.

2.1. Audio Processing

High-quality binaural audio samples are available in the datasets of the DCASE 2019 challenge. Therefore, binaural audio processing methods become applicable. We recommend the primary ambient extraction, which was originally proposed to upmix a stereo audio clip into arbitrary number of channels in order for them to be played back by any reproduction systems [16][17].

The primary ambient extraction assumes that in every time frame, there is a primary component and an ambient component in each channel. The primary components of two channels are assumed to be correlated. They are only different in the amplitude. The ambient components of two channels are assumed to have the same energy, but being uncorrelated. The algorithms of the primary ambient extraction are available in [18]. The primary ambient extraction explores the phase information of a binaural audio sample, differing from the previous methods in the DCASE challenge that simply abandoned all the phase information.

2.2. Acoustic Feature

Every audio sample is firstly resampled to 44100 Hz. Then, we use 2048-sample (46ms) Hanning windows and the hop-size of 1024-sample (23ms) to divide an audio sample into 431 frames. The spectrogram is generated by using the short time Fourier transform (STFT). After perceptual weightings are applied, the spectrogram is converted to the mel scale and passed through 128-bin mel filter bank. Finally, the mel spectrogram in the shape of (431,128, *N*) is obtained, where *N* denotes the number of channels. The aforementioned process has already been implemented in the hidden Markov toolkit (HTK) [19].

In this report, three different kinds of features are used. The 1-channel feature is obtained from the monaural audio, which is a mixture of the binaural audio. Although being lack of the spatial information, the monaural audio makes the non-spatial information more prominent. The 2-channel feature comes from the binaural audio. Because the mel spectrogram only counts for the magnitude, limited spatial information is embedded in the 2-channel feature. Furthermore, we obtain the 4-channel feature by using the primary ambient extraction. As mentioned above, the 4-channel feature provides more spatial information than the 2-channel feature.

2.3. Data Augmentation

A successful model contains abundant parameters and therefore massive data are required for training process of these parameters. However, the development dataset is not sufficiently large. Data augmentation is always carried out in some ways to avoid overfitting and enhance the model's generalization. We combine the cropping and mixup methods.

The cropping method generates more training data [20]. A feature with the size of (431,128, *N*) is cropped to 8 features with the size of (129,128, *N*) by using a hop-size of 43 samples. With this cropping method, we generate three additional cropped features based on every raw feature described in Section 2.2

The mixup method is a form of neighborhood risk minimization [21]. The mixup method fills up the gaps between the training samples, such that the model is improved to predict data not included in the training dataset. Interpolation of two features generates virtual features, while the labels are interpolated in the same way. This process is expressed as

$$\begin{aligned} \tilde{x} &= x_i + (1-\lambda)x_j \\ \tilde{y} &= y_i + (1-\lambda)y_j \end{aligned}, \qquad (1)$$

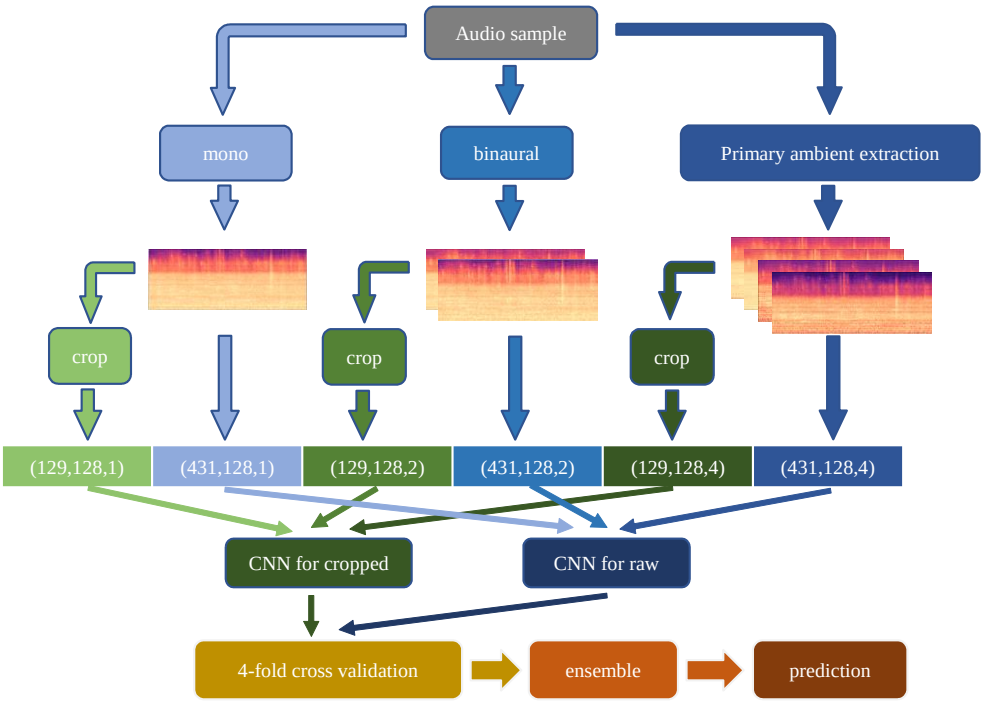

Figure 3: Overall architecture

where, xi and xj are two randomly chosen features; yi and yj are the corresponding labels. The random variable λ follows the beta distribution Be(α,α). When the hyper parameter α approaches zero, the regression of the mixup model will become the empirical risk minimization (ERM) [22].

3. NETWORK STRUCTURE

CNNs are able to recognize displacement, scaling and other distortion invariant feature maps. The layout of a CNN mimics the biological neural network. Due to its dedicated structure of local weight sharing, the CNN has distinguished advantages in speech recognition and image processing. Local weight sharing reduces not only the complexity in the structure of the CNN but also the complexity of data reconstruction in feature extraction and classification.

The structures of the CNNs that we implement are shown in Figures 1 and 2. Those models are inspired by the VGGNet [23]. Each model consists of 1 input layer, 8 convolution layers, 1 fully connected layer, and 1 output layer. The difference between the two models is that the first two convolution layers use different patch sizes for different input. Batch normalization (BN) is included in both the models to accelerate the learning process and improve the baseline level by regularization terms [24]. BN has been applied in most of the recent network architectures and been explained to be incompatible with dropout. Therefore, we have decided not to adopt the dropout.

3.1. Network Ensemble

Ensemble is a powerful method widely applied in various machine learning tasks. Ensemble of several weak models results in more reliable decisions. In our method, multiple CNN models are put together to achieve improved accuracies. The outputs of every CNN model after training are input to a random forest for final decision-making. The averaging method is also attempted for comparison. We have conducted a 4-fold cross validation on the development dataset. The overall architecture is shown in Figure 3.

4. EXPERIMENTS

4.1. Datasets

The TAU urban acoustic scene 2019 dataset consists of various acoustic scene samples collected in 6 cities in Europe. It has been extended from the TUT urban acoustic scene 2018 dataset by adding in another 6 cities in Europe. The binaural samples have been recorded for 5-6 minutes at different locations. The original recordings are divided into short segments of 10 seconds. Available information about the samples includes the acoustic scene, city name, and location where the recording was taken placed.

The dataset used in the training model is the development dataset of the TAU urban acoustic scenes 2019. The development dataset contains 40 hours of data from 10 cities, which have been divided into 14400 segments, i.e. 144 acoustic scene classes per

city. The training/test setup includes segments from Milan only to the test subset. There are 9185 segments in the training part, 4185 segments in the testing part and additional 1030 segments that were recorded in Milan.

4.2. Training Procedure

After audio preprocessing, the mel spectrogram features are normalized by the min-max method to avoid numerical problems and accelerate the convergence. The optimizer uses stochastic gradient descent (SGD) algorithm, whose learning rate, decay, and mini batch size are set to 0.01, 0.0001, and 32, respectively. Nesterov momentum at 0.9 is used to accelerate the SGD algorithm. For each model, we have conducted 4-fold cross validation and trained four times with data of different distributions to obtain stable and reliable results. The total number of models is counted by 3 (number of channls) × 2 (raw features + cropped features) × 4 (cross validation), which is 24 in total.

5. RESULTS

5.1. Results on the Kaggle Public Leaderboard

Table 1 presents our results on the Kaggle public leaderboard. So far, 6 models combined with the random forest ('6models_rf') has achieved the highest classification accuracy. According to the previous works, there is a risk of overfitting. We also propose to combine 6 models with the random forest and 4-fold cross validation ('6models-4folds_rf'), which is expected to be more stable and reliable. Moreover, the averaging of 6 models with the 4-fold cross validation ('6models-4folds_avg') achieves a considerable improvement as compared to the baseline. Furthermore, we present 4 models combined with the random forest and 4-fold cross validation ('4models-4folds_rf'), in which all the primary and ambient features are abandoned. This will help us to examine the effectiveness of the primary ambient extraction.

Table 1: Classification accuracies on the Kaggle public leaderboard

Method	Classification accuracy
6models_rf	0.843
6models-4folds_rf	0.840
4models-4folds_rf	0.833
6models-4folds_avg	0.830
Baseline	0.643

5.2. Submissions

Judging by our results on the Kaggle public leaderboard, we choose four models to submit at last:

1. **task1a_6_rf:** This submission is the output of 6 models combined with the random forest ('6models-rf').
2. **task1a_6-4folds_avg:** This submission is the averaging of 6 models with the 4-fold cross validation ('6models -4folds_avg').
3. **task1a_4-4folds_rf:** This submission is the output of 4 models combined with the random forest and 4-fold cross validation ('4models-4folds_rf').
4. **task1a_6-4folds_rf:** This submission is the output of 6 models combined with the random forest and 4-fold cross validation ('6models-4folds_rf').

6. CONCLUSIONS

In this paper, we have introduced the primary ambient extraction in the audio preprocessing stage of the acoustic scene classification, in order for the spatial information to be preserved. The results of the Kaggle public leaderboard demonstrate that proposed method improves the classification accuracy and the best reliable result achieves 0.84 at the time when this technical report is submitted.

7. REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, "Introduction to sound scene and event analysis," in *Computational Analysis of Sound Scenes and Events*, Springer, 2018.

[2] A. Mesaros, T. Heittola, and T. Virtanen, "DCASE 2017 challenge setup: tasks, datasets and baseline system," IEEE AASP Challenge on DCASE 2017 Technical Report, 2017.

[3] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M.D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.

[4] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," IEEE AASP Challenge on DCASE 2018 Technical Report, 2018.

[5] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," IEEE AASP Challenge on DCASE 2016 Technical Report, 2016.

[6] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification: an overview of DCASE 2017 challenge entries," in *16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, 2018.

[7] Y. Han and K. Lee, "Convolutional neural network with multiple-width frequency-delta data augmentation for acoustic scene classification," IEEE AASP Challenge on DCASE 2016 Technical Report, 2016.

[8] Y. Le Cun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed. Cambridge, MA: MIT Press, 1995, pp. 255–258.

[9] Y. Li, X. Li, and Y. Zhang, "Deep learning techniques for audio representation and classification," IEEE AASP Challenge on DCASE 2018 Technical Report, 2018.

[10] X. Xu, X. Chen, and D. Yang, "Acoustic scene classification using autoencoder," IEEE AASP Challenge on DCASE 2017 Technical Report, 2017.

[11] M. Dorfer, B. Lehner, H. Eghbal-zadeh, H. Christop, P. Fabian, and Widmer Gerhard, "Acoustic scene classification with fully convolutional neural networks and i-vectors," IEEE AASP Challenge on DCASE 2018 Technical Report, 2018.

[12] L. Yang, X. Chen, and L. Tao, “Acoustic Scene Classification Using Multi-Scale Features,” IEEE AASP Challenge on DCASE 2018 Technical Report, 2018.
[13] L. Zhang and J. Han, “Acoustic scene classification using multi-layered temporal pooling based on deep convolutional neural network,” IEEE AASP Challenge on DCASE 2018 Technical Report, 2018.
[14] Y. Han, J. Park, and K. Lee, “Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification,” IEEE AASP Challenge on DCASE 2017 Technical Report, 2017.
[15] 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.
[16] L. Chen, C. Shi, H. Li “Primary ambient extraction for random sign Hilbert filtering decorrelation,” in *the 23rd International Congress on Acoustics*, Aachen, Germany, September 2019;
[17] M. Goodwin, JM. Jot. “Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement,” in *the 2017 International Conference on Acoustic, Speech, and Signal Processing*, Honolulu, Hawaii, April 2007.
[18] J. He, *Spatial audio reproduction with primary ambient extraction*, Singapore: Springer Publishing Company, 2016.
[19] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book*, version 3.4. Cambridge University, March 2009.
[20] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” In *the 26th Annual Conference on Neural Information Processing Systems*, December 2012.
[21] H. Zhang, M. Cisse, Y. N. Dauphin,and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” in arXiv: 1710.09412, 2017.
[22] V. Vapnik and A. Y. Chervonenkis. “On the uniform convergence of relative frequencies of events to their probabilities,” in *Theory of Probability and its Applications*, 1971.
[23] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in arXiv:1409. 1556, 2014
[24] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *the 2015 International Conference on Machine Learning*, Lille, France, July 2015, pp. 448-456.