THE I2R SUBMISSION TO DCASE 2019 CHALLENGE

Teh Kah Kuan, Hanwu Sun and Tran Huy Dat

*Acoustic, Speech and Language Department, Institute for Infocomm Research, A*STAR Singapore*

ABSTRACT

This paper proposes Convolutional Neural Network (CNN) ensembles for acoustic scene classification of tasks1A of the DCASE 2019 challenge. In this approach various preprocessing features method: mel-filterbank and delta feature vectors, harmonic-percussive and subband power distribution are used to train CNN model. We also used score-fusion of the features to find an optimum feature configuration. On the official leaderboard data set of the task1A challenge, an accuracy of 79.67% is achieved.

Index Terms— Acoustic scene Classification, deep neural network, convolution neural network, network ensemble

1. INTRODUCTION

This paper describes the systems developed by Institute for Infocomm Research (I2R) team participating in the DCASE (Detection and Classification of Acoustic Scenes and Events) from Acoustic Scene Classification (Task1a) Challenge 2019 [1], which is designed to promote research in the area of acoustic scenes classification on single channel audio, under noisy conditions and distorted by the environment. The main objectives of this challenge include benchmark state-of-the art technology and support the development of new ideas and technologies in the area of acoustic scenes classification [1].

2. SYSTEM ARCHITECTURE

The proposed system is illustrated in Fig.1. The system is composed of 3 stages. First, the audio signal is converted to various time-frequency representations. These features are then fed to the CNNs for training the models. Finally, probability outputs of the CNN ensembles are used to produce the scene labels.

2.1. Audio Preprocessing

The DCASE 2019 data set is sampled at 48 kHz, we down sample to 44.1 kHz. Next, applying short-time Fourier transformation (STFT). The window function of STFT is a hann window, the window size is 2,048 samples, and the hop size is 1,024 samples. Finally, mel-spectrogram is obtained by applying 128 Mel bank. Mel spectrogram was converted to a logarithmic scale, normalized by dividing by the standard deviation and subtracting the mean value.

2.2. Harmonic-Percussive Source Separation

Mel-spectrogram is obtained from Harmonic-percussive source separation (HPSS), Han et al [2]. Librosa is used to separate an audio signal into its harmonic and percussive components.

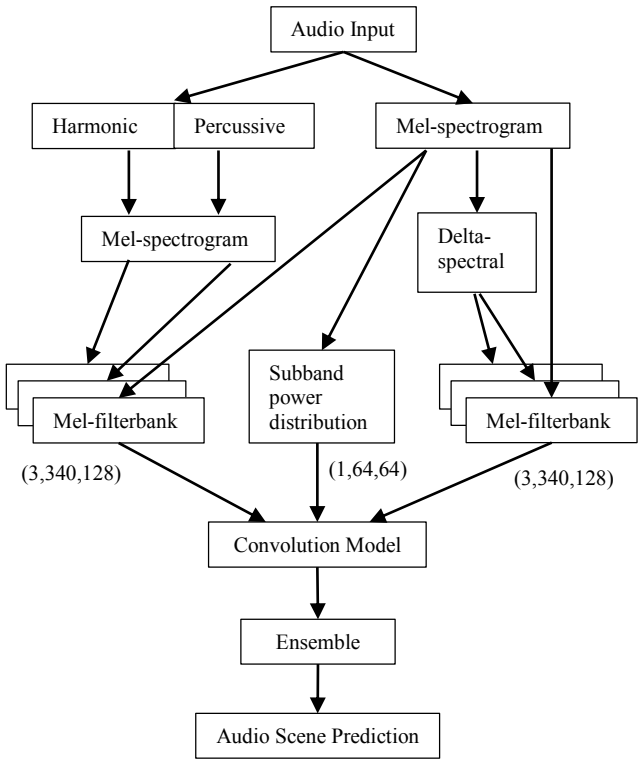

Figure 1: Architecture of the proposed system

2.3. Subband Power Distribution (SPD) Image Features

The SPD [3] captures the distribution of the log-spectral power of sound in each frequency subband over time and then stacked together to form a two dimensional image representation of frequency against normalized spectral power. A total of 64 bins are used with the bin edges equally spaced over the range of the normalized spectral power.

2.4. Network architecture

The network we used is CNN model proposed by Han et al [2]. In contrast, we combined the mel-spectrogram and calculated harmonic and percussive components of audio and create the 3 channel features (3, 430, 128). We also create another CNN model using stacked mel filter bank, delta filter bank and double delta filter bank to captures dynamic acoustic scene information.

3. EXPERIMENTS

3.1. Datasets

The dataset for this task is the TUT Urban Acoustic Scenes 2019 dataset, consisting of recordings from various acoustic scenes. The dataset was recorded in 12 large European cities, in different locations for each scene class. For each recording location there are 5-6 minutes of audio. The original recordings were split into segments with a length of 10 seconds that are provided in individual files. The dataset includes 10 scenes which are Airport, Indoor shopping mall, Metro station, Pedestrian street, Public square, Street with medium level of traffic, Travelling by a tram, Travelling by a bus, Travelling by an underground metro, Urban park. The dataset containing 40 hours of audio, balanced between classes.

3.2. Data Augmentation

Mixup [4] is a recent technique to improve generalization by increasing the support of the training distribution. In this technique, a new training sample by mixing a pair of two training samples. A new training sample $(\hat{x}, \hat{y})$ computed from the data and label pair (x_i, y_i), (x_j, y_j) by the following equation:

$$\hat{x} = \lambda x_i + (1-\lambda) x_j \quad (1)$$

$$\hat{y} = \lambda y_i + (1-\lambda) y_j \quad (2)$$

where $\lambda \sim beta(\alpha, \alpha)$. The hyperparameter λ controls the amount that is mixed in from the second samples.

3.3. Methods

Following methods have been implemented an evaluated.

- CNN_FB_HPSS – CNN with mel-filterbank + Harmonic-percussive features
- CNN_FB_HPSS_mixup - CNN_FB_HPSS and mixup training
- CNN_FB_delta - CNN with mel-filterbank + delta + delta-delta
- CNN_SPD – CNN with subband power distribution features

3.4. Networks Ensemble

In this section we ensemble all the models mentioned above and evaluate the performance on DCASE 2019 challenge task1A. The fusion is carried out in two ways: (i) weighted averaging voting and (ii) linear fusion using FoCal Multi-class toolkit [5]. We trained the system parameters on the developed dataset by optimizing the multi-class linear logistic regression cost. Let *M* be the number of classifiers, scores from all subsystems were combined with a linear combiner, as follows

$$S(t) = \sum_{k=1}^{M} a_k \times s(k,t) + \beta \quad (3)$$

where a_k are the weights, β is an *M*-dimensional bias vector (*M* = 3 in this case), and the *t* index indicates trials.

4. RESULTS

Table.1 shows the experimental results. Since model uses all data of Development dataset, it describes only accuracy of leaderboard dataset. In all results, it exceeds the accuracy of baseline system [6]. We have experimented with the choice of α in mixup, and found that values 0.2 worked well for our CNN architecture. Our experiments shows that 9-layes CNN worked best for this dataset.

Method	Accuracy
DCASE2019 Baseline	64.33
CNN_FB_HPSS (i)	76.17
CNN_FB_HPSS_mixup (ii)	76.33
CNN_FB_delta (iii)	NA
CNN_SPD (iv)	NA
Weighted averaging voting (ii+iii+iv)	79.17
FoCal Multi-class fusion (ii+iii_iv)	79.67

Table 1: Overall classification accuracies over methods

Finally, the performances of different ensemble methods are also presented. The fusion of system results is empirically important to improve the robustness of systems. It achieved 79.67% classification accuracies using FoCal Multi-class fusion.

5. CONCLUSIONS

In this paper, we have applied various preprocessing features method to CNN structure for task1A of DCASE2019 challenge. It significantly improves over the baseline system when combine together in an ensemble model. As a results, we could achieved 79.67 accuracy for task1A on the leaderboard dataset.

5. REFERENCES

[1] http://www.cs.tut.fi/sgn/arg/dcase2019/.

[2] 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.

[3] Pooja K J, Usha L, 2015, Robust Sound Event Recognition using Subband Power Distribution Image Feature, IJERT Volume 04, Issue 05, 2015.

[4] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup:Beyond Empirical Risk Minimization," in arXiv:1710.09412, 2017.

[5] N. Brummer, *FoCal Multi-class: Toolkit for Evaluation, Fusion and Calibration of Multi-class Recognition Scores*. [Online]. Available: http://niko.brummer.googlepages.com/focalmulticlass

[6] https://github.com/toni-heittola/dcase2019_task1_baseline/.