ACOUSTIC SCENE CLASSIFICATION BASED ON ENSEMBLE SYSTEM

Technical Report

Biyun Ding

Tianjin University
School of Electrical and
Information, 92 Weijin Road
Tianjin, 300072, China
1398491993@qq.com

Ganjun Liu

Tianjin University
School of Electrical and
Information, 92 Weijin Road
Tianjin, 300072, China
1364165995@qq.com

Jinhua Liang

Tianjin University
School of Electrical and
Information, 92 Weijin Road
Tianjin, 300072, China

ABSTRACT

This technical report is for the Task 1A Acoustic scene classification of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE). In this task, the features of audio will affect the performance. To improve the performance, we implement Acoustic scene classification task using multiple features and applying ensemble system which composed of CNN and GMM. According to the experiments which were performed with the DCASE 2019 challenge development dataset, the class average accuracy of GMM with 103 features is 64.3%, which is an improvement of 4.2% compared to Baseline CNN. Besides, the class average accuracy of the ensemble system is 66.3% , which is an improvement of 7.4% compared to Baseline CNN.

Index Terms— Acoustic scene classification, convolutional neural network, gaussian mixture model, ensemble system

1. INTRODUCTION

A lot of information is present sounds we perceive in our everyday environment and physical events that take place in it. With our natural ears we can recognize and classify individual sounds sources (rain, sub way coming, glass break, etc.). Developing signal processing methods to automatically extract this information has huge potential in several applications, for example searching for multimedia based on its audio content, and intelligent monitoring systems to recognize activities in their environments using acoustic information. However, a significant amount of research is still needed to reliably recognize sound scenes and individual sound sources in realistic soundscapes, where multiple sounds are present, often simultaneously, and distorted by the environment.

During the last decades, different methods have been proposed for the Acoustic scene classification such as the use of a bag-of frames approach that adopts a GMM (Gaussian mixture model) in combination with MFCC (Mel-frequency cepstral coefficients) features on relatively large time scales. This method has established itself in the field of SC (scene classification) and till today is still considered as a reasonable baseline system for the DCASE 2013 and DCASE 2016. DCASE 2019 challenge [1] task 1 is essentially an extended version of the previous DCASE 2013, 2016, 2017 and 2018 ASC task, providing a larger amount of data for an increased number of scenes.

In the acoustic scene classification task on DCASE 2013, DCASE 2016, DCASE2017, and DCASE2018 challenge, a number of novel approaches have been proposed [2]. In DCASE 2013, the baseline system for scene classification was based on "bag-of-frames" MFCC+GMM approach, that MFCC is mel-frequency cepstral coefficients and GMM is Gaussian mixture models. Most of the submissions used hand-made acoustic features along with classifier such as Hidden Markov Models (HMM), Support Vector Machines (SVM), and Random Forest. Some techniques that widely used for image processing such as a histogram of gradients (HOG) [3] and recurrence quantification analysis (RQA) [4] features also achieved top places. There was also an approach that utilises deep learning such as [5] using restricted Boltzmann machine, but it showed moderate classification accuracy, presumably due to small amounts of data. DCASE 2016 task 1 is essentially an extended version of the previous DCASE 2013 ASC task, providing a larger amount of data for an increased number of scenes. Many of submissions applied a deep learning approach such as a convolutional neural network (CNN) [6] and recurrent neural network (RNN) [7]. Although deep learning approach has been successful, top ranks were achieved by i-Vector and non-negative matrix factorization (NMF), which are rather conventional dictionary learning methods. Also, about half of submitted algorithms in this challenge used MFCC, one of the most popular hand-made features. [8] DCASE 2017 task 1 is essentially an extended version of the previous DCASE 2016 ASC task, providing a larger amount of data for a same number of scenes. Many of submissions applied CNN, Multilayer Perceptron (MLP), SVM, RNN, GMM and their fusion. And top ranks were achieved by CNN and the fusion with other classifiers such as MLP, SVM, and RNN. Also, most of submitted algorithms in this task used log-mel energies features and MFCC, which are the most popular hand-made features. Compared with DCASE 2016 ASC task, DCASE 2018 ASC task introduces a new dataset for acoustic scene classification. It has smaller number of classes than data from previous challenges, but it is much larger in size and acoustic variability, having been recorded in multiple cities across Europe [9]. Most of submissions applied CNN and their fusion. And top ranks were achieved by CNN and the fusion with other classifiers such as DNN, MLP, SVM, and GMM. Also, most of submitted algorithms in this task used log-mel energies features, which is the most popular hand-made features. It is worth to mention that the submitted ranked nine system applied

Xception as classifier and log-mel energies as the input. The system didn't use any data augment methods and channels except mono. Besides, most top ranks submissions applied muti-channels such as binaural, left, right and difference, also data augment methods such as mixup, block-mixing, and pitch-shifting.

As can be seen from the results of the DCASE task in the past, most system that based on CNN obtained good performance. Deep learning technology is rapidly evolving every day and one of the most important research topics in the audio processing field at the moment.

This report describes our submissions for Task 1A – Acoustic Scene Classification (ASC) in the DCASE-2019 Challenge. The basic approach to building our final classifier is based on GMM and CNN using multiple features. The following sections describe the details of the proposed system and the experimental results and conclusions.

2. SYSTEM FRAMEWORK

In this classification task, a segment of audio is classified into a single predefined class for single-label classification. The learning examples are audio segments with a single class annotated throughout. The annotations are encoded into target outputs which are used in the learning stage together with audio signals. In this case, classes are mutually exclusive. This condition is included into the neural network architecture by using output layer with softmax activation function, which will normalize outputted frame-level class presence probabilities to sum up to one. System block diagram of acoustic scene classification are shown in Fig. 1.

First, the datasets is split into disjoint training and testing sets. The training set is used to lead better-performing systems and the testing set is to provide more precise and reliable estimates of system performance. Then the features of training set is extracted and the single label corresponding to the audio segment is encoded into target outputs of acoustic scene classification model. The extracted feature set are applied to the training and testing stages.

In the training stage, the extracted features of the training set and the target outputs are input to the initial acoustic scene classification model. These are used to train the model and search for the optimal model that would separate the audio from different classes.

In the testing stage, we only need to extract the selected features in training stage, and input it to the acoustic scene classification model obtained through the training stage. Finally, the system performance is obtained by evaluating the outputs of testing stage.

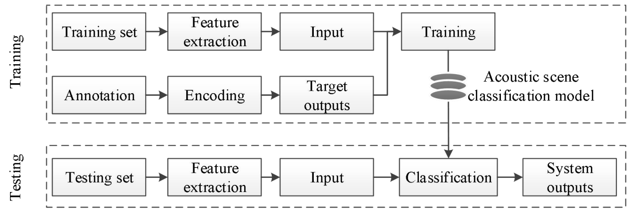

Figure 1: System block diagram of acoustic scene classification.

2.1. Feature extraction

The effectiveness of features determine the upper limits of the performance of the acoustic scene classification, and the classifier determines the extent to which performance approaches the upper limit. Therefore, feature extraction is vital importance in audio analysis of acoustic scene classification. In the audio analysis system, feature extraction can be utilized to transform the signal into a representation. It can represent the audio in a compact and non-redundant way requiring a small amount of memory and computational power.

Generally, the time domain features of a sound signal is not easy to interpret directly. It is nearly impossible to discriminate between sound scenes with most of the time domain features. Therefore, frequency-domain features and time-frequency domain features have been used to represent the sound signals that are more in line with the human perception [10].

Feature extraction incorporates a priori knowledge of acoustics, sound perception, or specific properties into an audio scene. The most common acoustic features are mel-band energies and MFCCs. They are based on the observation that human auditory perception focuses only on magnitudes of frequency components. The perception of these magnitudes is highly non-linear, in addition, perception of frequencies is also non-linear. Following perception, these acoustic feature extraction techniques use non-linear representation for magnitudes including power spectra and logarithm, and nonlinear frequency scaling such as mel-frequency scaling. The non-linear frequency scaling is implemented using filter banks which integrate the spectrum at non-linearly spaced frequency ranges, with narrow band-pass filters at low frequencies and with larger bandwidth at higher frequencies. Mel-band energies and MFCCs provide a compact and smooth representation of the local spectrum, but neglect temporal changes in the spectrum over time, which are also required for the recognition of environmental sounds. According to [10], log-mel energies features get a good performance in acoustic scene classification task.

Submission is based on log-mel energies, MFCC, first derivative of MFCC (D-MFCC), second derivative of MFCC (DD-MFCC), Zero crossing rate (ZRC), Root mean square energy (RMSE), and Spectrum centroid.

2.2. Classifiers

As described as section 1, we can see that the common classifiers for acoustic scene classification task include HMM, GMM, MLP, CNN, et al. GMM consists of a weighted mixture of K multivariate Gaussian distributions. HMM and GMM are widely used in audio applications. MLP is a forward-structured artificial neural network that maps a set of input vectors to a set of output vectors. It is the simplest and oldest "deep" model. And CNN is a class of deep, feed-forward artificial neural networks, most commonly applied to analyzing visual imagery. In baseline system of Task1A, the CNN structure is shown in figure 2. There are 2 convolution layers and max-pooling in CNN. Maximum pooling is performed after each convolution layer. Zero padding is added before the convolution layer to make the most of the edge. There are 32 filters in the first layer and 64 filters in the second layer, and the size of convolution kernel is 3x3 each convolution filters. All convolution layers and pooling layers are with a stride of 1. The dropout layer is with the rate of 0.3 except for the last one.

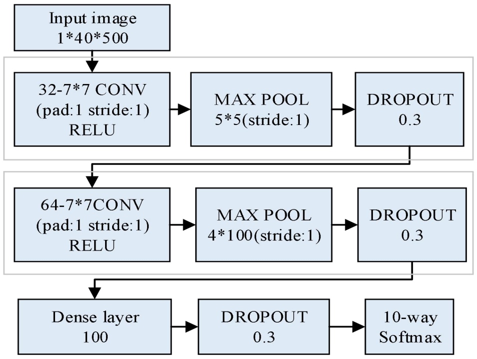

Figure 2: The CNN structure in baseline system of Task1A

In addition to the CNN of the baseline system, we also use the classifier based on GMM, MLP and Xception.

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1. Datasets

For performance assessment, DCASE2019 dataset consists of recordings from 10 acoustic scenes, including airport, bus, metro, metro_station, park, public_square, street_pedestrian, street_pedestrian, shopping_mall, tram, was used. The baseline data set is important in the comparison algorithm and in the study of the reproduction of results under various conditions. The experimental data set is from the dataset of DCASE2019, extending the TUT Urban Acoustic Scenes 2018 dataset with other 6 cities to a total of 12 large European cities. A total 1440 segments (2400 minutes of audio), recorded at 48 kHz with 24-bit resolution in stereo, were provided per scene and the length of the audio segments were 10 seconds. The dataset size is increased compare to 2018, but the length of each audio segment is same as 2018.

3.2. Features

In this work, different features are used in single and multichannel modes. All features are extracted from audio signals. We used the features in two modes, single-channel and 5-channels. In single channel mode, the audio signal is first converted to mono and single-channel features are extracted from it. In the 5-channels mode, five sets of features are extracted from the signal from 5-channels. Two feature sets from left (L) and right (R) channels, one from the subtraction of both channels (i.e. S = L−R), and the last two feature sets respectively from the Harmonic and Percussive audio separated from mono channel via Harmonic-percussive source separation (HPSS). For HPSS, librosa [11] which is a Python package for music and audio analysis is used, and initial values are used for parameters. We use the features mergered by these 5 feature sets as a single input to the classifier. Here, the classifier tries to use all channels at the same time to use all the available information.

The features commonly applied in acoustic scene classification task include log-mel energies and MFCC. In our work, we mainly use log-mel energies features.For extracting these features, first short time Fourier transform is computed on 40 ms Hamming windowed frames with 20 ms overlap using 2048 point FFT, and then , the spectrograms are obtained. Next, the spectrograms is transformed to 40 or 128 Mel-scale band energies, finally, log of these energies is taken. Therefore, a log-mel energies feature vector of size 40 × 500 is obtained from each audio clip of 10 second.

The second set of features is obtained as 20-dimensional MFCC from spectrograms. In addition to these features above, first derivative of MFCC (D-MFCC), second derivative of MFCC (DD-MFCC), Zero crossing rate (ZRC), Root mean square energy (RMSE), and Spectrum centroid are also used in acoustic scene classification task to represent audio signal.

3.3. Development system results

These results on the development set is shown in Table 1. And they are based on log-mel energies including 40 dimensions features. We compare the system performance among the baseline CNN, the Minimal MLP and GMM based method on the development dataset. From these result, the performance of GMM based is better than baseline system, but MLP based system conversely worse than baseline system.

Table 1: Average scene accuracy for the baseline CNN, the Minimal MLP and GMM based method on the development dataset.

Scene	Accuracy (%)		
	Baseline CNN	Minimal-MLP	GMM
airport	43	53.7	45.6
bus	61	83.4	73.5
metro	65.4	54.3	59.8
metro_station	52	28.3	49.2
park	85	88.6	82.4
public_square	42.4	35.9	45.5
street_pedestrian	55.6	50.6	57.6
street_traffic	57.1	35.4	66.2
shopping_mall	84.6	64.7	86.6
tram	71.6	25.2	56
Overall	61.7	52	62.2

Besides, as shown as table 2, we calculated the validation accuracy for the GMM with different features on the development dataset. For the features, LM and MF respectively donate 40-dimensional log-mel energies and 20-dimensional MFCC features. LM+MF donates 60-dimensional features merged by LM and MF. AF donates all 103-dimensional features including 40-dimensional log-mel energies, 20-dimensional MFCC, 20-dimensional D-MFCC, 20-dimensional DD-MFCC, 1-dimensional ZRC, 1-dimensional RMSE, and 1-dimensional Spectrum centroid. GL and GL2 respectively donate 128-dimensional log-mel energies features with 48kHz and 44.1kHz sample rate of audio. Finally 5C donates 240-dimensional features merged five 40-dimensional log-mel energies extracted from the signal from 5-channels.

As shown as table 2, the performance of log-mel energies features are similar to MFCC features in GMM based system. However, GMM based system combing log-mel energies and MFCC features obtains better performance than the single type features. moreover, GMM based system combing seven type features totaled 103-dimensional obtains the best performance in all the GMM based systems. It means the more features and the number of feature type, the better GMM based system performance is.

Table 2: Validation accuracy for the GMM with different features on the development dataset.

Scene	Accuracy (%)						
	LM	MF	LM +MF	AF	GL	GL2	5C
airport	45.6	44.7	40.9	43.2	45.8	43.5	55.1
bus	73.5	71.6	71.6	73.7	77.8	76.6	68.7
metro	59.8	58	61.9	67.9	64	53.3	55
metro_station	49.2	45.5	44.4	47.8	47.8	47.4	54
park	82.4	83.7	82.1	83.7	77.5	80.8	80.1
public_square	45.5	48.6	47.5	50.4	53	61	50.4
street_pedestrian	57.6	71.4	68	60.5	52.4	53.5	50.1
street_traffic	66.2	60.1	65.3	68.3	65.7	57.1	69.5
shopping_mall	86.6	84.8	85.8	89.3	88.8	89.9	89.6
tram	56	53.7	57.8	57.8	58.9	62.6	58.5
Overall	62.2	62.2	62.5	**64.3**	63.2	62.6	63.1

During addressing the ASC task, we attempted nearly 20 systems to improve the system performance. The parameter settings of these systems in DCASE 2019 task1a is shown in table 3. Where N donates the index of individual system. The sampling rate and feature dimension is respectively expressed by fs and dim. For the CNN applied in baseline system include two convolution layers, so it is abbreviated into 2-CNN. For the urgent challenge time, we didn't test the development system performance of Xception based system.

To obtain better system performance, we tried some ensemble system based these system. In other words, I use these systems as subsystems of ensemble system to improve system performance.

According to experiment, we obtain an ensemble system integrating the results of 16 systems by majority vote. The indexes of these 16 systems are number 1 to 16. The overall test accuracy is 65.7% .

Besides, We found the testing performance of system 15 is better than other systems. To maximize the performance of ensemble system, based on 16 systems above, we increase the weight of system 15 in ensemble system via using it twice. The overall test accuracy of new ensemble system is 66.3% . The confusion matrix is shown in figure 2.

Table 3: The parameter settings of systems in DCASE 2019 task1a.

N	system	fs(kHz)	dim	classifier	Accuracy(%)
1	baseline	48	40	2-CNN	61.7
2	v1_baseline_fs_12k	12	40	2-CNN	60
3	v1_baseline_fs_12k_500	12	40	2-CNN	58.4
4	v1_baseline_fs_24k	24	40	2-CNN	61.3
5	v1_baseline_fs_24k_500	24	40	2-CNN	60.3
6	v5_MFCC	48	40	2-CNN	59.1
7	v5_logmel+MFCC	48	60	2-CNN	59.7
8	v5_all_features	48	103	2-CNN	61.7
9	v3_GMM	48	40	GMM	62.2
10	v4_GMM_all_features	48	103	GMM	**64.3**
11	v4_GMM_logmel+MFCC	48	60	GMM	62.5
12	v4_GMM_MFCC	48	20	GMM	62.2
13	Mini-MLP	48	40	Mini-MLP	52
14	Mini-MLP_200	48	40	Mini-MLP	50.9
15	v2_GMM	48	128	GMM	**63.2**
16	v2_GMM_44.1	44.1	128	GMM	62.6
17	v11_GMM_muti-channel	48	240	GMM	63.1
18	v7_Xception	48	40	Xception	

According to experiment, we obtain an ensemble system integrating the results of 16 systems by majority vote. The indexes of these 16 systems are number 1 to 16. The overall test accuracy is 65.7% .

Besides, We found the testing performance of system 15 is better than other systems. To maximize the performance of ensemble system, based on 16 systems above, we increase the weight of system 15 in ensemble system via using it twice. The overall test accuracy of new ensemble system is 66.3% . The confusion matrix is shown in figure 3.

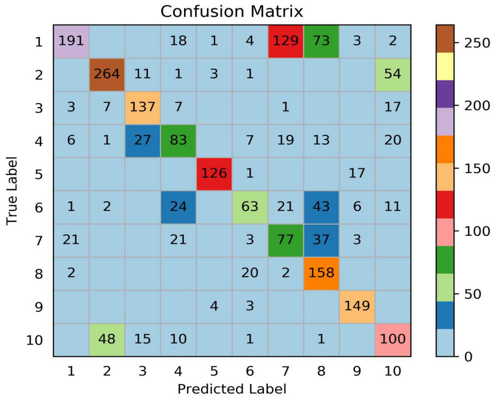

Figure 3: The confusion matrix of new ensemble system

3.4. DCASE 2019 Submission

For the final submission, we submitted a result for task1A following the challenge rule. We submit the system that based on log-mel energies or multiple features and ensemble system. We submitted four system result including (1)the v2_GMM, (2) ensemble system including 13 subsystems, (3) v4_GMM_all_features, and (4) ensemble system including 14 subsystems. Here, he indexes of the second and last submitted systems respectively are [1, 2, 3, 5, 6, 9, 10, 11, 12, 15, 16,17] and [1, 2, 3, 5, 6, 9, 10, 11, 12, 15, 15, 16,17].

4. CONCLUSION

In this report, we focused on exploring the application of CNN and GMM for acoustic scene classification (Task 1a). We found that log-mel energies are better than others features in 2-CNN based systems, but log-mel energies are similar to MFCC features in GMM based system. And the performance of simple GMM based classifier better than 2-CNN in some case. Besides, we found that the validation of features are not only related to the characters of task but also the type of classifier. 2-CNN isn’t always effective for ASC task. To improve the performance of the system, it is necessary to adjust the network structure and parameter settings of the CNN. Although we attempted to use 4-CNN with four convolution layers and 8-CNN with eight convolution layers in ASC task, but the lager parameters lead the training very slow, so we didn't get the result of 4-CNN and 8-CNN. In addition, the model is trained by inputting appropriate features.

5. REFERENCES

[1] http://dcase.community/challenge2019/.

[2] A. Mesaros, T. Heittola, and T. Virtanen, “Tut database for acoustic scene classification and sound event detection,” in *Signal Processing Conference (EUSIPCO), 2016 24th European. IEEE*, 2016, pp. 1128-1132.

[3] Bisot, Victor, S. Essid, and G. Richard. "HOG and subband power distribution image features for acoustic scene classification." *Signal Processing Conference IEEE*, 2015:719-723.

[4] G. Roma, W. Nogueira, P. Herrera, and R. de Boronat, “Recurrence quantification analysis features for auditory scene classification,” *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, vol. 2, 2013.

[5] J. Nam, Z. Hyung, and K. Lee, “Acoustic scene classification using sparse feature learning and selective max-pooling by event detection,” *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[6] Y. Han and K. Lee, “Convolutional neural network with multiple-width frequency-delta data augmentation for acoustic scene classification,” *DCASE2016 Challenge*, Tech. Rep., September 2016.

[7] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, and B. Schuller, “The up system for the 2016 DCASE challenge using deep recurrent neural network and multiscale kernel subspace learning,” *DCASE2016 Challenge*, Tech. Rep., September 2016.

[8] Yoonchang Han, Jeongsoo Park and Kyogu Lee, “Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification,” *DCASE2017 Challenge*, Tech. Rep., September 2017.

[9] Mesaros, Annamaria , T. Heittola , and T. Virtanen . “A multi-device dataset for urban acoustic scene classification.” , 2018.

[10] Virtanen, Tuomas, M. D. Plumbley, and D. Ellis, “Computational Analysis of Sound Scenes and Events,” *Springer International Publishing*, 2018.

[11] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in python,” in *Proceedings of the 14th python in science conference*, 2015, pp. 18-25.