LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION USING SIMPLE CNN

Technical Report

Biyun Ding

Tianjin University
School of Electrical and Information, 92 Weijin Road
Tianjin, 300072, China
beatrice@tju.edu.cn

ABSTRACT

This technical report describes our Acoustic Scene Classification systems for DCASE2021 challenge Task1A: Low-Complexity Acoustic Scene Classification with Multiple Devices. In this work, many factors affect the performance. To improve the performance while ensure the model complexity, we attempt different methods in term of features, sampling rate, channel, classifier type, the network architecture of CNN, and the post-processing of predictions. According to the experiments on TAU urban acoustic scenes 2020 mobile development dataset, the best accuracy of single system we implemented is 55.89%, which is an improvement of 7% compared to Baseline CNN. Besides, the accuracy of the late fusion is 59.80% , which is an improvement of 11.35% compared to Baseline CNN.

Index Terms— Acoustic scene classification, convolutional neural network, gaussian mixture model, late fusion.

1. INTRODUCTION

Acoustic scene classification (ASC) is a classification task of assigning predefined semantic labels to audio streams recorded in a certain environment by analyzing audio signals [1]. The semantic labels describe the environment information of the audio streams. Developing signal processing methods to automatically extract the environment related 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.

During the last decades, many research have been done to reliably recognize sound scenes and individual sound sources in realistic soundscapes. For example, the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) is conducted since 2013 to stimulate research in the ASC field. The DCASE challenge provides the benchmark data and a competitive platform to promote sound scene research and analyses. DCASE 2021 challenge task 1 [2] is essentially an extended version of the previous DCASE 2013, 2016, 2017, 2018, 2019, and 2020 ASC task, providing an increased number of data for different scenes and more factors.

In the ASC task results of the previous DCASE challenges, a number of novel approaches have been proposed [3] for the Acoustic scene classification. 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. Also, most of submitted algorithms in ASC tasks used Log mel-band energies features, which is the most popular hand-made features. It is worth to mention that 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.

In the case of subtask A, the audio files in the dataset are identical to the previous year, but only change in model complexity. The main issue of this task is to design a model under 128KB. This corresponds to 32,768 non-zero parameter when converted to a 32-bit floating-point per parameter. It is very small number considering last year's participants submitted more than millions or billions of parameters.

This report describes our submissions for Task 1A – Acoustic Scene Classification (ASC) in the DCASE-2021 Challenge. The following sections include details of our model structure and training methods. Due to the model size limitation in subtask A, it neccessary to simple the model based on neural network. The basic approach to building our final classifier is based on GMM and CNN using Log mel-band energies as 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. The system block diagram of acoustic scene classification are shown in Fig. 1.

In this framework, 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. In the training stage, the

acoustic features of training set is extracted as the input of model and the single label corresponding to the audio segment is encoded into target outputs. The input and the target outputs are inputted train the acoustic scene classification model. In the testing stage, we extract the same features as training stage and input them into the well-trained acoustic scene classification model to get the prediction results. Finally, the system performance is obtained by evaluating the outputs of testing stage.

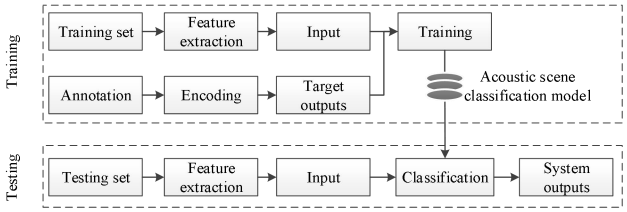

Figure 1: System block diagram of acoustic scene classification.

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 [4].

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 Mel Frequency Cepstral Coefficients (MFCC). They are based on the observation that human auditory perception focuses only on magnitudes of frequency components. Mel-band energies and MFCC 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 [4], Log mel-band energies features get a good performance in acoustic scene classification task.

In this work, we employed log-mel band energies (logMel), MFCC, first derivative of MFCC (ΔMFCC), second derivative of MFCC (ΔΔMFCC), Zero crossing rate (ZRC), Root mean square energy (RMSE), and Spectrum centroid, Linear Prediction Coefficients (PLP), Constant Q Transform (CQT), and Linear Prediction Coefficients (LPCC) features.

2.2. Classifiers

For classifier, we employed CNN, GMM, BLS, and SVM to model the acoustic scenes. Where CNN is a class of deep, feed-forward artificial neural networks, most commonly applied to analyzing visual imagery. In the baseline system of Task1A, the CNN structure is shown in Table 1. 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 16 filters in the first layer and 32 filters in the second layer, and the size of convolution kernel is 7x7 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.

Table 1: The CNN structure in baseline system of DCASE 2021 Task1A.

Architecture	Parameters
Input layer	40 * 500 (10 seconds)
layer #1	2D Conv (16×7), Batch normalization, ReLu activation
layer #2	2D Conv (16×7), Batch normalization, ReLu activation, max pooling (5, 5) + Dropout (30%)
layer #3	2D Conv (32×7), Batch normalization, ReLu activation, max pooling (4, 100) + Dropout (30%)
Flatten	
Dense layer #1	Dense layer (100, ReLu), Dropout (30%)
Output layer	10-way softmax
Learning: 200 epochs (batch size 16), data shuffling between epochs Optimizer: Adam (learning rate 0.001)	

Noted that: the model size of baseline system is 90KB, and its accuracy 47.7% (± 0.9). In this task, a model complexity limit of 128 KB is set for the non-zero parameters [5].

2.3. Post-processing

Commonly, the system output could be divided into single output and fusion output. The single output is obtained by only one classifier. The fusion output is obtained by fusing the predictions of multiple classifiers, which can obtain a tremendous boost in classification accuracy. The method that fuses multiple systems to obtain the final prediction is called as late fusion, which is generally used to take advantage of different systems and thus improve classification performance,

The late fusion can stabilize and generalize the final results. The commonly used late fusion techniques include SVM, regression, voting methods like the average, majority, and weighted voting, where the voting method is comparatively simple and effective. We applied late fusion while fusing different ASC model scores combined to classify the acoustic scenes well.

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1. Datasets

The ASC task of the DCASE 2021 challenge continues to use TAU Urban Acoustic Scenes 2020 Mobile. It contains data from 10 cities and 9 devices: 3 real devices (A, B, C) and 6 simulated devices (S1-S6). Data from devices B, C, and S1-S6 consists of randomly selected segments from the simultaneous recordings. Therefore all overlap with the data from device A, but not necessarily with each other.

This dataset includes ten classes (three indoor, four outdoor, and three transportation), consisting 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. A total 23040 segments (64 hours of audio), recorded at 48 kHz with 24-bit resolution, were provided per scene and the length of the audio segments were 10 seconds. The organizer of the challenge provides basic metadata of training/test split consisting of 13,965 samples in the training set and 2,970 samples in the test set. The dataset size is increased compare to 2019, but the length of each audio segment is same as 2019.

3.2. Features

In this work, different features are used. All features are extracted from audio signals. We used the features in two modes, single-channel and multi-channel. In single channel mode, the audio signal is first converted to mono and single-channel features are extracted from it. For multi-channel audio, we firstly generate the Harmonic and Percussive audio separated from mono channel via Harmonic-percussive source separation (HPSS). It implemented by librosa [6], which is a Python package for music and audio analysis is used, and initial values are used for parameters.

The features commonly applied in acoustic scene classification task include logMel and MFCC. In our work, we mainly use logMel 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 logMel feature vector of size 40×500 is obtained from each audio clip of 10 second.

In addition to logMel features, we also employ MFCCs, ZRC, RMSE, Spectrum centroid, PLP, and LPCC features for acoustic scene classification task.

3.3. Classifier

Table 2: The CNN architecture of different 2-CNN versions for DCASE 2021 Task1A.

2-CNN version	layer #1	layer #2	layer #3	Flatten layer
cnn2d_2 cnn2d_8	16×7	16×7, pool(5×5)	32×7, pool(4×100)	GlobalMaxPooling2D Flatten, Dense(100)
cnn2d_3 cnn2d_12	16×5	16×5, pool(5×5)	32×5, pool(4×100)	GlobalMaxPooling2D Flatten, Dense(100)
cnn2d_4 cnn2d_13	16×3	16×3, pool(5×5)	32×3, pool(4×100)	GlobalMaxPooling2D Flatten, Dense(100)
cnn2d_9 cnn2d_16	16×3	16×3, pool(3×3)	32×3, pool(4×100)	GlobalMaxPooling2D Flatten, Dense(100)
cnn2d_5 cnn2d_14	8×7	8×7, pool(5×5)	16×7, pool(4×100)	GlobalMaxPooling2D Flatten, Dense(100)
cnn2d_6 cnn2d_15	8×5	8×5, pool(5×5)	16×5, pool(4×100)	GlobalMaxPooling2D Flatten, Dense(100)
cnn2d_7 cnn2d_10	8×3	8×3, pool(5×5)	16×3, pool(4×100)	GlobalMaxPooling2D Flatten, Dense(100)
cnn2d_11 cnn2d_17	8×3	8×3, pool(3×3)	16×3, pool(4×100)	GlobalMaxPooling2D Flatten, Dense(100)

In this work, we apply CNN, BLS, SVM, and GMM based classifier. When calculate their model size, we found that the model size of some simple CNN-based classifiers are less than 128KB. The model size of both BLS-based and SVM-based classifier are more than 128KB. The model size of all GMM-based classifiers are less than 128KB. In a word, CNN-based and GMM-based classifier can satisfied the requirement of model complexity.

To ensure the model complexity of CNN-based, we attempt multiple simpler CNN versions than the baseline system. Their architecture is reported in Table 2. Here, cnn_2d_8 is the baseline system of DCASE2021 task1 A.

3.4. Developed systems

To improve the performance while ensure the model complexity, we attempt different methods in term of features, sampling rate, channel, classifier type, the network architecture of CNN, and the post-processing of predictions.

Table 3: The factors of the acoustic scene classification system.

Factors	Range
Sampling rate	44.1kHz, 22.05kHz
Features	LogMel, MFCCs, CQTs, ZRC, RMSE, PLP, LPC, Spectrum centroid
Channel	Mono, H, P[1]
Classifier	CNN, GMM
Network architecture of CNN	2-CNN, 3-CNN, 4-CNN, 5-CNN[2]
Post-processing	Single output, late fusion output

1: H and P channel are the Harmonic and Percussive audio separated from mono channel via HPSS, respectively.

2: 2-CNN is the CNN-base classifier with 2 convolutional layers. Similarly, 3-CNN, 4-CNN, and 5-CNN has 3, 4, and 5 convolutional layers, respectively.

3.5. System results

These results on the development set is shown in Table 3. And they are based on logMel including 40 dimensions features. We compare the system performance with different classifier on the development dataset. From these result, it can be seen that CNN-based classifier outperforms others under the same factors.

Table 3: The accuracy of ASC system with CNN, SVM, BLS and GMM on the development dataset.

Scene	Accuracy (%)			
	CNN_2_8 (Baseline)	SVM	BLS	GMM
airport	18.86%	28.96%	30.98%	29.97%
bus	42.76%	40.74%	28.62%	26.60%
metro	44.11%	44.44%	44.11%	46.13%
metro_station	29.97%	28.28%	26.26%	28.28%
park	80.81%	64.98%	51.18%	50.84%
public_square	42.09%	25.93%	20.88%	23.57%
street_pedestrian	69.70%	46.46%	48.48%	36.70%
street_traffic	41.08%	21.55%	19.87%	31.65%
shopping_mall	58.59%	66.67%	63.97%	62.96%
tram	56.57%	26.60%	31.99%	44.44%
Overall	**48.45%**	39.46%	36.63%	38.11%

Noted: the system employed 40 dimensions logMel features. The sampling rate is the default value 44.1 kHz.

For the features, the dimension of logMel is 40 and mfcc is 20. all_feat donates all 103-dimensional features including 40-dimensional logMel, 20-dimensional MFCC, 20-dimensional Δ MFCC, 20-dimensional Δ Δ MFCC, 1-dimensional ZRC, 1-dimensional RMSE, and 1-dimensional Spectrum centroid.

As shown as table 4, the CQTs feature outperforms other features when the ASC system bases on GMM. The performance of logMel features are similar to MFCC features in GMM based system. PLP and LPCC features perform poorly. Moreover, GMM based system combing seven type features totaled 103-dimensional obtains the best performance in all the GMM based systems. It means that GMM based system combing multiple features obtains better performance than the single type features.

Table 4: The accuracy of ASC system basing on GMM along with different features on the development dataset.

Feature	Logmel	mfcc	all_feat	plp	lpcc	CQTs
Accuracy %	38.11	37.74	39.73	34.48	23.94	48.32

Table 5: The parameter settings of systems in DCASE 2021 task1a.

N	fs (kHz)	Feature	dim	classifier	Log loss	Accuracy	Model size
1	44.1	logMel	256	cnn2d_12	1.519	55.89 %	116.07KB
2	44.1	logMel	134	cnn2d_10	1.360	55.45%	78.58KB
3	44.1	logMel	134	cnn2d_8	1.523	54.58%	115.33 KB
4	44.1	logMel	134	cnn2d_12	1.440	53.00%	78.57 KB
5	44.1	logMel	256	cnn2d_13	1.707	52.96%	91.57KB
6	44.1	logMel	134	cnn2d_2	1.327	52.22%	76.30 KB
7	44.1	logMel	256	cnn2d_3	1.263	51.45%	39.55 KB
8	44.1	logMel	134	cnn2d_13	1.485	51.25%	54.07 KB
9	44.1	logMel	134	cnn3d_1	1.310	50.07%	69.41 KB
10	44.1	logMel	134	cnn2d_3	1.268	50.00%	39.55 KB
11	44.1	logMel	40	cnn2d_2	1.619	49.87%	76.30 KB
12	44.1	logMel	134	cnn3d_2	1.336	48.72%	36.04 KB
13	44.1	logMel	134	cnn2d_5	1.350	48.55%	19.79 KB
14	**44.1**	**logMel**	**40**	**Baseline**	**1.892**	**48.45%**	**90.33 KB**
15	44.1	CQTs	-	GMM	2.303	48.32%	640 B
16	44.1	logMel	40	cnn2d_10	1.658	48.11%	53.58 KB
17	44.1	logMel	134	cnn2d_15	1.469	47.95%	31 KB
18	44.1	logMel	40	cnn2d_5	1.486	46.00%	19.79KB

It is noted that the model parameters of CNN above are quantized to float16 after training.

According to experiment, the best accuracy of single system we implemented is 55.89%, which is an improvement of 7% compared to Baseline CNN.

During addressing the ASC task, we attempted nearly 150 systems to improve the system performance. To obtain better system performance, we tried some fusion system based these system. In other words, we use these systems as subsystems of fusion system to improve system performance.

To find the best subset size in fusion system, we perform a experiment shown in Table 6. In the experiment, we obtain several fusion systems fusing the results of multiple systems by majority vote, here the subset size is set from 3-12.

From the results we can see that the highest accuracy reach 59.80%, which is higher than the accuracy of the best single system (accuracy: 55.89% referring to Table 6). It meas that the late fusion method is effective to improve the system performance. The best fusion system achieves an improvement of 11.35% compared to Baseline CNN.

Table 6: Performance of the late fusion method at the system subset with the different sizes selected from the system set.

Subset size	3	4	5	6	7
Accuracy	58.82%	58.92	**59.80%**	59.49%	59.70%
Subset size	8	9	10	11	12
Accuracy	59.46%	59.76%	58.89%	58.69%	58.79%

Noted: Subset size means the number of systems used to be fused to obtain the fusion output.

As we know, the system for DCASE2021 task1a is ranked by macro-average multiclass cross-entropy (Log loss). However, the log loss value of the fusion system with highest accuracy (59.80%) is 1.448, which is much higher than some single system. Because the accuracy is not completely consistent with the log loss value. So we employ the log loss value as the optimization objective of the fusion system. Then, we get a lower log loss 1.193 with 55% accuracy.

3.6. DCASE 2021 Submission

For the final submission, we submitted a result for task1A following the challenge rule. We submit the system that based on logMel and fusion system. We submitted four systems. We submitted four system result including:

(1) **Ding_TJU_task1a_1** : system 2, 1.360, 55.45%, 78.58KB;
(2) **Ding_TJU_task1a_2** : system 7, 1.263, 51.45%, 39.55 KB;
(3) **Ding_TJU_task1a_3** : system [8,13,17,18], 1.193, 55.0%, 124.64KB;
(4) **Ding_TJU_task1a_4** : system 10, 1.268, 50.00%,39.55 KB.

4. CONCLUSION

In this report, we focused on exploring the application of CNN and GMM for acoustic scene classification (Task 1a). We found that logMel are better than others features in 2-CNN based systems. To improve the classification performance and satisfied the requirement of model complexity (limitation of 128KB), it is necessary to adjust the network structure and parameter settings of the CNN. Although we attempted to use 8-CNN with eight convolution layers in ASC task (achieves 57.58% accuracy), but the lager parameters can not satisfied the requirement of model complexity, so we just employ simper models.

5. REFERENCES

[1] Waldekar S, Saha G. Two-level fusion-based acoustic scene classification. *Appl Acoust* 2020;170:. https://doi.org/10.1016/j.apacoust.2020.107502107502.

[2] DCASE2021 website: http://dcase.community/challenge2021/task-acoustic-scene-classification.

[3] 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.

[4] Virtanen, Tuomas, M. D. Plumbley, and D. Ellis, "Computational Analysis of Sound Scenes and Events," *Springer International Publishing*, 2018.
[5] The baseline system of DCASE2021 task1a : http://dcase.community/challenge2021/task-acoustic-scene-classification#baseline-systems.
[6] 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.