# ACOUSTIC SCENE CLASSIFICATION BASED ON FHR_MOBILENET

Technical Report

*Hongxia Dong, Lin Zhang, Xichang Cai, Menglong Wu, Ziling Qiao, Yanggang Gan, Juan Wu*

North China University of Technology, Beijing, China
caixc20_ncut@126.com

## ABSTRACT

This technical report describes our submission for Task1 of DCASE2022 challenge. We calculated 128 log-mel energies under the original sampling rate of 44.1KHz for each time slice by taking 2048 FFT points with 50% overlap. Additionally, deltas and delta-deltas were calculated from the log Mel spectrogram and stacked into the channel axis. The resulting spectrograms were of size 128 frequency bins, 43 time samples and 3 channels with each representing log-mel spectrograms, its delta features and its delta-delta features respectively. Then, the three channel feature map is fed into the mobilenet-based frequency high-resolution network. Finally, after $1 \times 1$ convolution and global average pooling, the classification results are obtained through softmax output. The classification accuracy of our proposed model is 53.9% with a loss value of 1.378. The number of parameters of the model is 70.608K, where each parameter is represented using int8 and the MACs are 28.461M.

***Index Terms—*** Acoustic Scene Classification, Data Augmentation, FHR_Mobbilenet

## 1. INTRODUCTION

Acoustic Scene Classification (ASC) is a task of classifying given data to a place where it was recorded. Each data corresponds to one class out of ten, and there is no data with multiple labels. The length of the data is one second, but the useful information appears very rarely. This task is one of the major topics that has been covered every year in the DCASE challenge. This year, several changes in the acoustic scene classification task have brought new research questions into focus. For example, limiting the number of parameters and the number of multiplicative accumulation operations of the model [1].

The main issue of the Task1 is to design a classifier that works stably on various microphone types. However, the development dataset mostly includes the data collected from a specific microphone, and the evaluation data will include data recorded with a microphone that has not appeared in the development set.

The following sections include details of our model structure and training methods. The complexity of the computation is mainly measured in terms of the number of parameters and MMAC (Million Multiplicative Product Operations-Memory Accesses). The size of the number of parameters needs to be limited to 128KB, and each parameter needs to be represented in 8-bit, and the number of parameters is including zero-valued parameters. The difference between this and DCASE 2021 is that last year each parameter can be represented in 8-bit or 32-bit, and the size of the MAC was not required last year, but this year it is limited to 30 MMAC.

## 2. AUDIO DATASET

The development dataset for this task is TAU Urban Acoustic Scenes 2022 Mobile, development dataset. The dataset contains recordings from 12 European cities in 10 different acoustic scenes using 4 different devices. Additionally, synthetic data for 11 mobile devices was created based on the original recordings. Of the 12 cities, two are present only in the evaluation set. The dataset has exactly the same content as TAU Urban Acoustic Scenes 2022 Mobile, development dataset, but the audio files have a length of 1 second (therefore there are 10 times more files than in the 2020 version).

Recordings were made using four devices that captured audio simultaneously. The main recording device consists in a Soundman OKM II Klassik/studio A3, electret binaural microphone and a Zoom F8 audio recorder using 48kHz sampling rate and 24-bit resolution, referred to as device A. The other devices are commonly available customer devices: device B is a Samsung Galaxy S7, device C is aniPhone SE, and device D is a GoPro Hero5 Session.

Audio data was recorded in Amsterdam, Barcelona, Helsinki, Lisbon, London, Lyon, Madrid, Milan, Prague, Paris, Stockholm and Vienna. The dataset was collected by Tampere University of Technology between 05/2018 - 11/2018. The data collection received funding from the European Research Council, grant agreement 637422 EVERYSOUND.

## 3. SYSTEM ARCHITECTURE

### 3.1. data preprocessing

The data of Task1 are mono audio files with 44.1 kHz sample rate. We transformed them into power spectrogram by skipping every 1024 samples with 2048 length Hann window. A spectrum of 51 frames was yielded from 1 seconds audio file, and each spectrum was compressed into 128 bins of Mel frequency scale. Additionally, deltas and delta-deltas were calculated from the log Mel spectrogram and stacked into the channel axis. The number of frames of the input feature is cropped by the length of the delta-delta channel so that the final shape becomes $[128 \times 43 \times 3]$.

### 3.2. data augmentation

We used the data augmentation means of mixup [4] and specaugment [5] to increase the diversity of data distribution. $\alpha$ parameter in mixup was set to 0.4, frequency_masking_para in specaugment was set to 13, and time_masking_para was set to 4. We did not use additional training datasets other than the official training dataset.

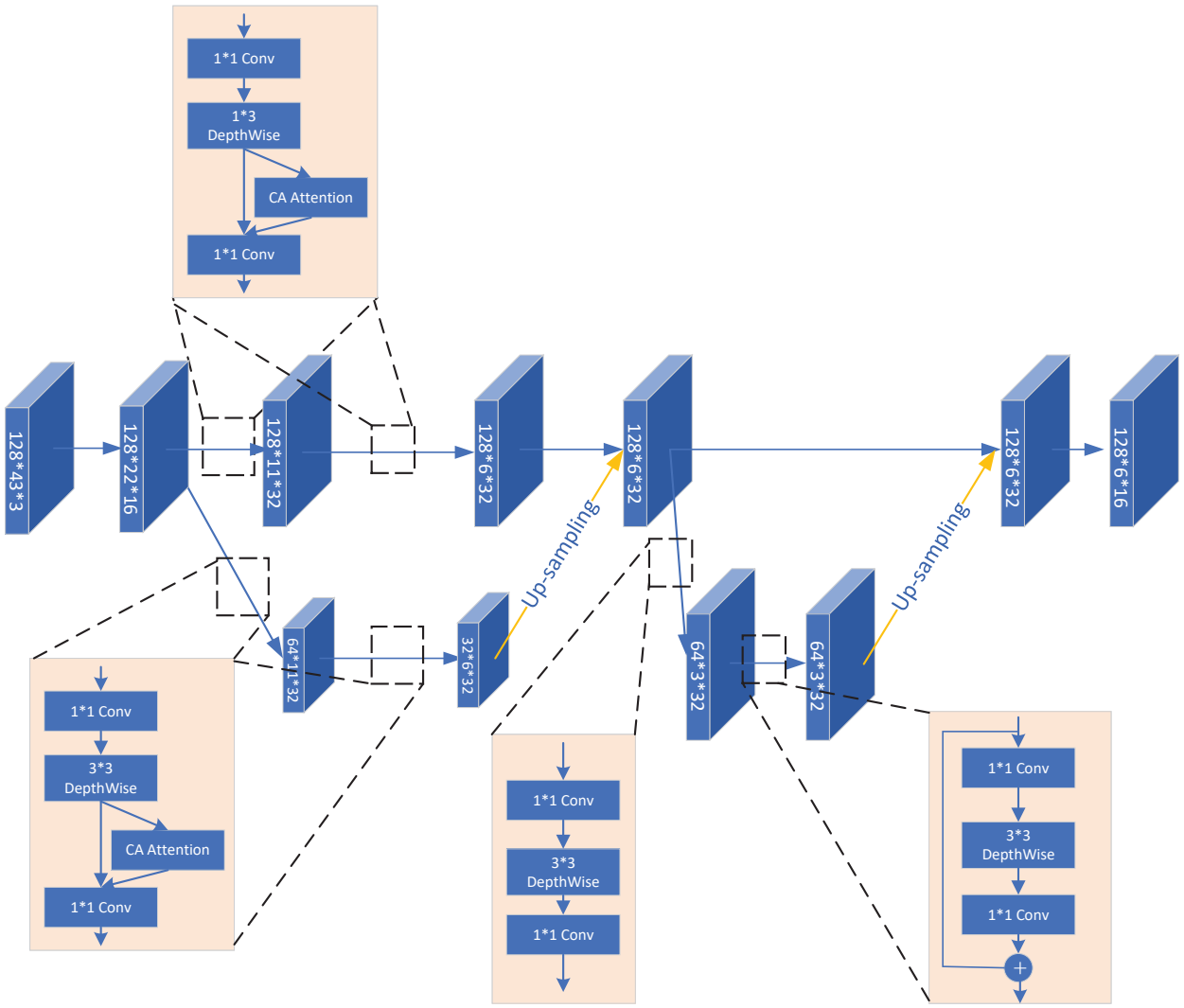


Figure 1: The overall structure of our model.

### 3.3. model design

The overall framework of our proposed model is shown in Figure 1, which adopts the idea of frequency high resolution, while applying the inverse residual module in the model. Let the feature mapping always maintain high resolution on the frequency axis and constant convolutional downsampling on the time axis.

The input feature map is first convolved by $1 \times 5$ with step size (1, 2) to obtain a $128 \times 22 \times 16$ feature map, and then passed through two inverse residual modules with coordinate attention mechanism, where the depthwise convolution has a convolution kernel size of $1 \times 3$ and step size of 1, and the feature map size is $128 \times 6 \times 32$; the other depthwise convolution has a convolution kernel size of $3 \times 3$ and step size of 2, and the feature map size is $32 \times 6 \times 32$. The feature map size is $32 \times 6 \times 32$, and the feature map size is $128 \times 6 \times 32$ by up-sampling the fea-ture map and adding the feature map of the first road. The output feature map is $64 \times 3 \times 32$, which is upsampled and the two maps are summed. Finally, after 16 $1 \times 1$ convolution kernels with a step size of 1, the size of the obtained feature map is $128 \times 6 \times 16$.

In addition, we conducted comparison experiments with the shufflenet_shallow model, and since the shufflenet model proposed in [7] exceeds the entry requirements in terms of the number of parameters and the number of MACs, we scaled down the number of repetitions of the shufflenet_block and adjusted the number of channels per stage to (120, 48, 96, 192, 100). And since the five-dimensional transpose operation in channelshuffle in Shufflenet v2 is not supported in the tflite model, the channelshuffle operation in the shuffle module is removed directly. All other settings are consistent with the method proposed in this paper. The model structure is shown in Table 2.

Table 1: Accuracy on the fold 1 evaluation set(class-wise)

| Scene label | Baseline | Shufflenet_shallow | Our system_3060Ti | Our system_2080Ti |
|---|---|---|---|---|
| Airport | 39.4% | 44.7% | 46.3% | 46.8% |
| Bus | 29.3% | 49.3% | 63.1% | 73.1% |
| Metro | 47.9% | 46.5% | 45.5% | 50.8% |
| Metro_station | 36.0% | 28.3% | 47.6% | 50.3% |
| Park | 58.9% | 63.4% | 59.7% | 76.2% |
| Public_square | 20.8% | 22.6% | 38.7% | 33.0% |
| Shopping_mall | 51.4% | 40.1% | 50.0% | 51.3% |
| Street_pedestrian | 30.1% | 36.0% | 34.1% | 40.1% |
| Street_traffic | 70.6% | 68.9% | 70.2% | 71.5% |
| Tram | 44.6% | 34.9% | 59.4% | 45.4% |
| **Average** | 42.9% | 43.5% | 51.5% | 53.9% |
| **Loss** | 1.575 | 2.132 | 1.424 | 1.378 |
| **Model_size** | $46.5K$ | $74.648K$ | $70.608K$ | $70.608K$ |
| **MACs** | $29.23M$ | $7.434M$ | $28.461M$ | $28.461M$ |

Table 2: Architecture of shufflenet_shallow in DCASE 2022

| Layer | Output shape | Kernel size |
|---|---|---|
| Input | $128 \times 43 \times 3$ | - |
| Conv2d1 | $64 \times 22 \times 120$ | $3 \times 3$ |
| MaxPool2d | $32 \times 11 \times 120$ | $3 \times 3$ |
| shuffle_block_s2 | $16 \times 6 \times 48$ | $3 \times 3$ |
| shuffle_block_s1 | $16 \times 6 \times 48$ | $3 \times 3$ |
| shuffle_block_s2 | $8 \times 3 \times 96$ | $3 \times 3$ |
| shuffle_block_s2 | $4 \times 2 \times 192$ | $3 \times 3$ |
| Conv2d2 | $4 \times 2 \times 100$ | $1 \times 1$ |
| GlobalAvg | $1 \times 100$ | - |
| Dense | $1 \times 10$ | $100 \times 10$ |
| Softmax | $1 \times 10$ | - |

### 3.4. Categorical Focal Loss

Focal loss attenuates the logloss generated by welltrained samples, so that the model can focus on the poorly trained samples. The following equation describes focal loss with balancing parameter $\alpha$, focusing parameter $\gamma$ and prediction score $p_t$,

$$FL\left(p_t\right) = -\alpha\left(1-p_t\right)^{\gamma}\log\left(p_t\right) \tag{1}$$

Increasing the value of $\gamma$ increases the sensitivity of the model to misclassified samples, and scales the loss function linearly. Our setting was $\gamma = 1.0$ and $\alpha = 0.3$, respectively.

### 3.5. Training Setup

We trained our model using Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9. We use warm up and cosine annealing to plan the learning rate. Choosing warmup learning rate can make the learning rate of several epochs or some steps smaller. Under the warmup learning rate, the model can gradually become stable. After the model is relatively stable, we choose the preset learning rate for training to make the convergence speed of the model faster, The effect of the model is better. The preset learning rate is $0.001$, warmup_learning is set to $4\mathrm{e}-06$. The epoch is set to 256 and batch_size is set to 128.

## 4. EXPERIMENTAL RESULT

The baseline system implements a convolutional neural network (CNN) based approach using log mel-band energies extracted for each 1-second signal. The network consists of three CNN layers and one fully connected layer to assign scene labels to the audio signals. The system is based on the DCASE 2021 Subtask A baseline system. Model size of the baseline when using TFLite quantization is 46512 parameters, and the MACS count is 29.23 M. The loss of baseline system was 1.575, and the classification accuracy was 42.9%. For our system, we used all the development data to train the model on 2080Ti and 3060 graphics computer respectively, and tested the evaluation set on 2080Ti computer using shufflenet_shallow model, the experimental results are shown in Table 1. The model size of our system after quantization with TFLite on a 2080Ti graphics computer is 70608 parameters with a MACS count of 28.461 M. The loss of the model is 1.378 and the classification accuracy is 53.9%. The loss of our system on a 3060 graphics computer is 1.424 and the classification accuracy is 51.5%. The model size of shufflenet_shallow quantized with TFLite is 74648 parameters, the MACS count is 7.434M, the loss of this model is 2.132, and the classification accuracy is 43.5%.

## 5. CONCLUSION

In this technical report, we proposed a acoustic scene classification system. We use log-mel spectrograms, deltas and delta-deltas and mobilenet-based frequency high-resolution network to improve the performance of the system. We achieved a classification accuracy of 53.9%, which is 11.0% over than the baseline system. Model size when using TFLite quantization is 70608 parameters, and the MACS count is 28.461 M.

## 6. REFERENCES

[1] http://dcase.community/challenge2022/.

[2] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.

[3] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5693-5703.

[4] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez -Paz, "Mixup: Beyond empirical risk minimization," arXiv preprint arXiv:1710.09412, 2017.

[5] Park D S, Chan W, Zhang Y, et al. Specaugment: A simple data augmentation method for automatic speech recognition[J]. arXiv preprint arXiv:1904.08779, 2019.

[6] Seo S, Kim J. Mobilenet using coordinate attention and fusions for low-complexity acoustic scene classification with multiple devices[J]. Tech. Rep., DCASE2021 Challenge, 2021.

[7] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 116-131.