LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION USING CONVOLUTION NEURAL NETWORK

Technical Report

Ee-Leng Tan, Jin Jie Yeo, Santi Peksi, Woon-Seng Gan

Centre of Information Sciences and Systems, Nanyang Technological University,
50 Nanyang Avenue, Singapore 639798
etanel@ntu.edu.sg, ye0024an@e.ntu.edu.sg, speksi@ntu.edu.sg, ewsgan@ntu.edu.sg

ABSTRACT

In this technical report, we describe the CISS-NTU team's submission for Task 1 Low-Complexity Acoustic Scene Classification of the detection and classification of acoustic scenes and events (DCASE) 2023 challenge [1]. We have explored and adapted the hyperparameters of the baseline (BL) system provided in this challenge. The TAU Urban Acoustic Scene 2022 Mobile, development dataset [2] has been used to train and validate our models. Each audio sample is transformed into 160 log-mel energies. Three models are submitted with two trained using the development dataset and one trained using the development dataset combined with augmented samples. The best performing model achieves an accuracy of 52.1% and a log loss of 1.372, and only requires 6.46 M of multiply-and-accumulate (MAC) operations and has a memory usage of 54.30 KB.

Index Terms— Acoustic scene analysis, CNN, data augmentation, mel spectrogram.

1. INTRODUCTION

In task 1 of DCASE challenge 2023, acoustic scene classification (ASC) is employed to recognize 10 acoustic scenes from 12 cities based on 1 sec audio samples. To align ASC with the performance of typical edge devices, task 1 [1] of the DCASE challenge 2023 has imposed the following system complexity requirements:

- Maximum memory allowance: 128KB
- Maximum number of MACS per inference: 30 MMAC

Mel-spectrogram from the audio signals is used as the input features for our models. The models' parameters and architecture were tuned to achieve the best performance within the above-mentioned complexity limits.

Convolution neural networks (CNNs) have dominated the entries of the task 1 challenge. Many CNN models have produced promising results on ASC tasks and achieved good accuracies on the TAU urban acoustic scene 2022 mobile dataset [2]. In this work, we adapt and optimize the CNN architecture of the baseline (BL) system, and the input features obtained from the audio samples in the development dataset. These modifications focus on improving classification accuracy while reducing system complexity. Augmentation techniques are experimented on and will be introduced to enhance the generalizability of the model to unseen devices and the variability of the audio samples. To further reduce the model size and computational cost, post-training quantization is applied to convert the weights and parameters of the trained model.

This report is organized as follows. In section 2, the input features, augmentation techniques used, and proposed model are discussed. Section 3 presents the results of our submissions based on the development and augmented datasets. This report is concluded in Section 4.

2. PROPOSED SYSTEM

2.1. Preprocessing

The TAU urban acoustic scene 2022 mobile dataset contains recordings of 10 acoustic scenes in 12 European cities. These recordings are captured using four devices and synthetic data for 11 devices was generated using the recordings. Each 1 sec audio sample is captured with a sampling frequency of 44.1 kHz sampling rate and encoded at 24-bit resolution.

For the input feature, a 160 bin mel-spectrogram is calculated using the short time Fourier transform (STFT) using a window length of 0.16 sec with 50% overlapping. This setting results in 13 frames and an input feature shape of [160 × 13].

44.1 kHz, 32 kHz, 16 kHz, and 8 kHz were experimented with to determine the optimal sampling frequencies. The input features and the downsampling of the audio samples is performed using Librosa [3]. A summary of the log loss and average accuracies (based on a modified BL model) across the 10 scenes is provided in Table I. We observed a very slight improvement when the sampling rate of the audio samples is at 44.1 kHz, as compared to the other sampling frequencies. A comparison of the averaged mel spectrograms of the 10 acoustic scenes is shown in Fig. 1. The acoustic scenes of the airport, park, shopping mall, street pedestrian, and tram are relatively narrower in terms of bandwidth, and we can see significant frequency components at 16kHz and above for the remaining acoustic scenes. Based on these observations, the sampling rate of the audio samples is selected to be 44.1 kHz.

2.2. Data Augmentation

To prevent overfitting and improve system robustness, several augmentation techniques using additive Gaussian noise, pitch-shift, time-shift, and SpecAugment are experimented with during

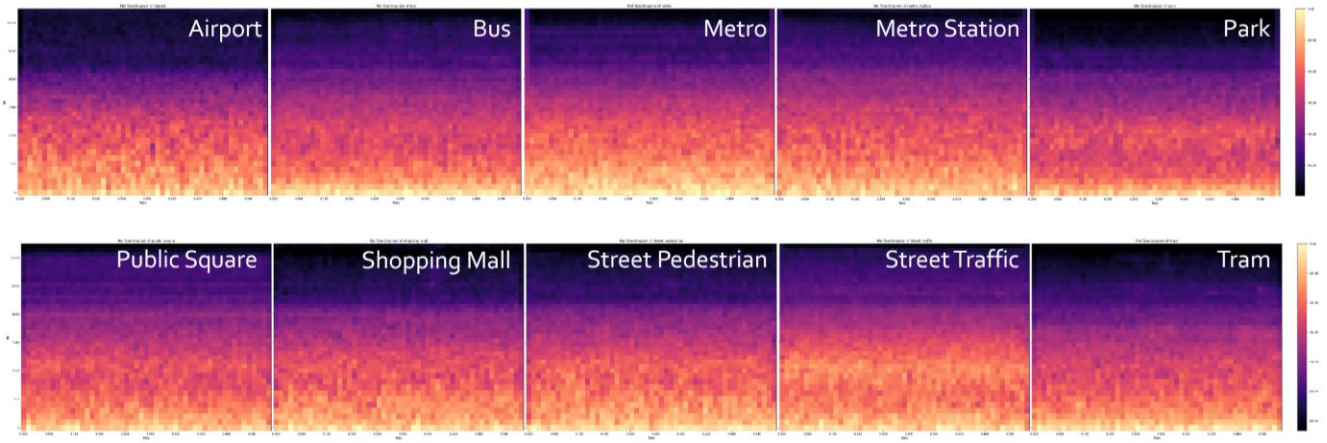

Fig. 1. Averaged mel-spectrograms of the 10 acoustic scenes. Acoustic scene of bus, metro, metro station, public square, and street traffic are found to span across wider bandwidth.

Table I Average Log Loss and Accuracies With Different Sampling Frequencies

Sampling Frequency (kHz)	Log Loss	Accuracy (%)
8	1.429	48.1
16	1.42	47.8
32	1.426	47.8
44.1	1.415	49.1

Table II Hyperparameters of models

Model	CNN Layer #1			CNN Layer #2			CNN Layer #3			D_{1_1}	MAC (M)	MEM (KB)	Aug
	Filters	K_{1_1}, K_{1_2}	M_{1_1}, M_{1_2}	Filters	K_{2_1}, K_{2_2}	M_{2_1}, M_{2_2}	Filters	K_{3_1}, K_{3_2}	M_{3_1}, M_{3_2}				
BL	16	7 × 7	N.A.	16	7 × 7	5, 5	32	7 × 7	4, 10	100	29.23	46.51	No
SM_1	16	3 × 3	2, 2	16	3 × 3	4, 2	32	7 × 7	6, 1	32	2.96	37.50	No
SM_2	16	3 × 3	2, 2	24	5 × 5	4, 3	32	7 × 7	6, 1	32	6.46	54.30	No
SM_3	16	3 × 3	2, 2	24	5 × 5	4, 3	32	7 × 7	6, 1	32	6.46	54.30	Yes

Table III Submited Models

Model	Overall Log Loss	Overall Accuracy (%)	Training (kWh)	Inference (kWh)
BL	1.542	41.2	0.1955	0.4594
SM_1	1.397	50.3	0.2692	0.0518
SM_2	1.372	52.1	0.4272	0.1061
SM_3	1.381	50.0	0.5516	0.1057

the training of the submitted models. The description of the four augmentations are as follows.

2.2.1. Gaussian Noise

Gaussian noise with an amplitude of 0.01 is added to the audio samples in the time domain. This augmentation technique helps to simulate the natural variation and noise that is found in real-world recording. By adding Gaussian noise to the audio samples, the model learns to be more robust against small variations and perturbations in the audio samples [4].

2.2.2. Pitch-Shift

Pitch-shifting involves changing the pitch of a waveform while maintaining the tempo of the audio sample. A shift of four semi-tones was found to be beneficial to environmental sound classification [5] and is applied in our audio samples.

2.2.3. Time-Shift

Audio samples are shifted forward and backward with rollover to generate more temporal variations of the audio samples. This augmentation exposes the model to a broader range of temporal patterns of the audio samples.

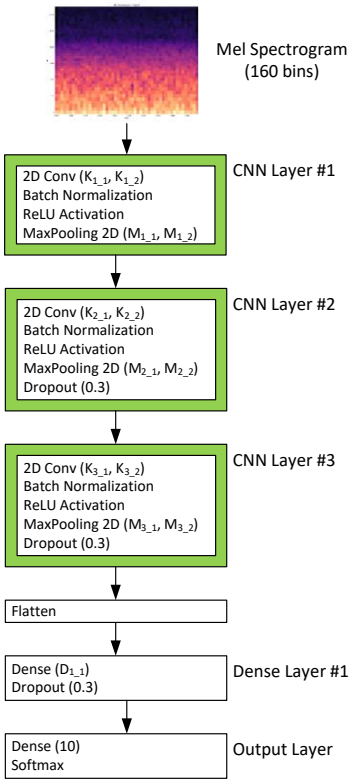

Fig. 2. Network architecture of submitted model (SM) adapted from BL system.

2.2.4. SpecAugment

SpecAugment [6] masks a set of frequencies in a spectrogram and generates variation in both the frequency and temporal structure of the audio samples. This technique is a popular augmentation technique in the 2022 challenge. For comparison, the power consumption of all models will be discussed in the next section.

2.2.5. Choosing Augmentation

Based on our experimental results, our model achieve the best performance with time-shift augmentation. To reduce power consumption, only one augmentation is applied in the training of our model.

2.3. Network Architecture

The BL model is used as the starting point of our development. Taking into consideration of the parameters and MAC limits of the challenge, several modifications were introduced to reduce the computational cost while improving the performance of the model.

The network architecture of the submitted models is illustrated in Fig. 2, and the hyperparameters used in these models are summarized in Table II. The dropout of dense layer #1 is increased from 0.3 to 0.5. Three variations of the models are submitted, with the SM_1 and SM_3 having the lowest and highest computational cost of 2.96 MMAC and 6.46 MMAC, respectively. The training of the two submitted models without augmentation incurred the lowest power consumption.

2.4. Post-Training Quantization

Post-training integer quantization is implemented using Tensorflow [7]. While the input and output of the models are kept at float32, the weights of the models are converted to int8. After the quantization, our submitted models have MMACs lower than 6.5M and the memory usage is kept below 55KB.

3. RESULTS AND SUBMISSION

The submitted models were trained for 450 epochs with a batch size of 64 using SGD optimizer and callbacks providing learning rate schedule and early stopping. Training of the models is terminated before the 450 epochs. The development dataset is split into training and validation sets at 70% and 30%, respectively.

The results of the submitted models are summarized in Table II. Model SM_3 is trained using the development dataset combined with augmented audio samples, and models SM_1 and SM_2 are trained using the development dataset only. Both SM_1 and SM_2 have accuracies over 50%, even though the system complexity is significantly lower than the BL system. In addition, SM_1 has a smaller memory footprint as compared to the BL system.

4. CONCLUSIONS

In this technical report, we described the CISS-NTU submissions to task 1 of the DCASE 2023 challenge. The proposed model is extended from the BL system after a series of optimization of the hyperparameters, focusing on the reduction of the system complexity while improving the accuracy of the model. Mel spectrograms are used as the input to the model, and time-shift augmentation is found to work best with the submitted models. Post-training quantization was applied to the trained model and the weights of the model were converted to int8. Our model achieves an accuracy of 52.1% with a memory usage of 54.30 KB, while only requiring a computational cost of 6.46 MMAC.

5. ACKNOWLEDGEMENT

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (MOE-T2EP20221-0014).

6. REFERENCES

[1] Irene Martín-Morató, Francesco Paissan, Alberto Ancilotto, Toni Heittola, Annamaria Mesaros, Elisabetta Farella, Alessio Brutti, and Tuomas Virtanen. Low-complexity acoustic scene classification in DCASE 2022 challenge, 2022.

[2] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. Acoustic scene classification in DCASE 2020 challenge: generalization across devices and low complexity solutions. In Proceedings of the Detection and Classification

of Acoustic Scenes and Events 2020 Workshop (DCASE2020), 56–60. 2020.

[3] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," SciPy, 2015.

[4] VV. Eklund, "Data Augmentation Techniques for Robust Audio Analysis," Master Thesis, Tampere University, Sept. 2019

[5] J. Salamon, and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279-283, Mar. 2017.

[6] D. S. Park, et al., "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," Interspeech, pp. 2613-2617, 2019.

[7] https://www.tensorflow.org/lite/performance/post_training_quant

[8] R. Balestriero, L. Bottou, and Y. LeCun, "The effects of Regularization and Data Augmentation are Class Dependent," 36th Conference on Neural Information Processing Systems, 2022.