ACOUSTIC SCENE CLASSIFICATION USING CONVOLUTION NEURAL NETWORK WITH LIMITED DATA

Technical Report

Ee-Leng Tan[1], Jisheng Bai[2], Jun Wei Yeo[1], Santi Peksi[1], Woon-Seng Gan[1]

[1] Smart Nation TRANS Lab, Nanyang Technological University,
50 Nanyang Avenue, Singapore 639798
[2] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
etanel@ntu.edu.sg, baijs@mail.nwpu.edu.cn, junwei004@e.ntu.edu.sg, speksi@ntu.edu.sg,
ewsgan@ntu.edu.sg

ABSTRACT

In this technical report, we present the SNTL-NTU team's submission for Task 1 Low-Complexity Acoustic Scene Classification of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2024 challenge [1]. The proposed CNN model is trained using the TAU Urban Acoustic Scene 2022 Mobile development dataset [2]. For the input of the model, each audio sample is transformed into log-mel energies. The model has a memory usage of 88.2 KB and requires 25.9M multiply-and-accumulate (MAC) operations. Using the development dataset, the proposed model achieved an accuracy of 46.02%, 49.54%, 53.89%, 56.43%, and 59.08% for 5%, 10%, 25%, 50%, and 100% of the development dataset, respectively.

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

1. INTRODUCTION

In Task 1 of the DCASE Challenge 2024, acoustic scene classification (ASC) is employed to classify 10 acoustic scenes from 12 cities based on 1-second 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 constraints:

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

To reduce model complexity, mel-spectrograms computed from audio signals are used as the input features for our models. The parameters and architecture of the proposed model were tuned to achieve the best performance within the above-mentioned complexity limits. In addition, five development splits are provided by the organizer of this challenge, where 100%, 50%, 25%, 10%, and 5% of the original development dataset can be used to train our models.

Convolutional neural networks (CNNs) have dominated the entries of the Task 1 challenge. Numerous 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 have adopted a model based on CNN. Augmentation techniques were experimented with and introduced to enhance the generalizability of the model to unseen devices and the variability of the audio samples obtained from different cities. 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 various splits of the development dataset. 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 input featurse, a 192 bin mel-spectrogram is calculated using the short time Fourier transform (STFT) using a window length of 0.18 sec with a 17% overlap. This confugration produces 33 frames and an input feature shape of [192 × 33] for each audio sample.

The input features and the down-sampling of the audio samples are computed and performed using Librosa [3]. The sampling rate 44.1kHz is selected based on the averaged mel spectrograms of the 10 acoustic scenes, as shown in Fig. 1. The acoustic scenes of the airport, park, shopping mall, street pedestrian, and tram are narrower in terms of bandwidth, while significant frequency components at 16kHz and above are observed for the remaining acoustic scenes.

2.2. Data Augmentation

Three augmentations using SpecAugment [4], Freq-MixStyle [5], and device impulse response (DIR) [6] are applied in proposed models to prevent overfitting and improve system robustness.

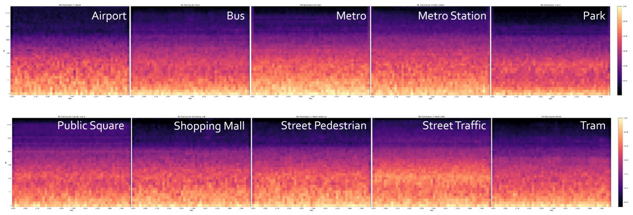

Fig. 1. Averaged mel-spectrograms of 10 acoustic scenes. Acoustic scenes of bus, metro, metro station, public square, and street traffic are found to span across a wider bandwidth.

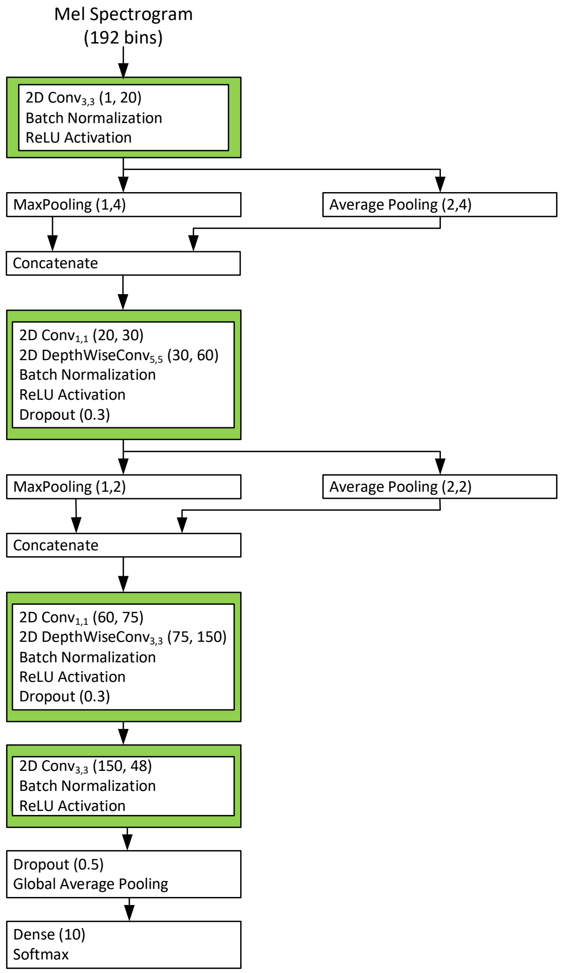

Fig. 2. Proposed model. Numbers in (,) indicate input and output channels, and kernel size of conv layers are denoted as subscripts.

SpecAugment typically consists of three types of augmentations: time warping, frequency masking, and time-masking. Frequency masking has proven particularly effective. Freq-MixStyle extends MixStyle [7] to the frequency domain of the audio samples. By exposing the model to a variety of mixed spectral properties, it generalizes to different acoustic environments and their variations. Research has demonstrated that models trained with Freq-MixStyle exhibits better generalization to unseen conditions, enhancing domain-invariance. Models trained with DIR augmentation can also better handle variations in recording devices, leading to improved accuracies with the development dataset.

2.3. Proposed Model

The network architecture of the submitted model is illustrated in Fig. 2. To minimize computational cost, the proposed model utilizes depthwise and pointwise convolutions. A combination of max-pooling and average-pooling is used to enhance the features at the earlier layers of the model.

Post-training dynamic quantization is implemented [8] to reduce the memory requirements of the proposed model. While the input and output of the models are kept at float32, the weights of the models are converted to INT8. After the quantization, the memory requirement of the submitted model is kept at 88.2KB.

3. RESULTS AND SUBMISSION

For all splits, the proposed model was trained for 150 epochs with a batch size of 256 using ADAM optimizer with learning rate scheduler. The results of the provided baseline and submitted model are summarized in Tables I and II, respectively. Compared to the baseline model, the proposed model demonstrates better classification across most classes, except for the street pedestrian class. Furthermore, this comparison highlights that the proposed model has improved comparatively more in the lower splits.

4. CONCLUSIONS

In this technical report, we described the SNTL-NTU submissions to task 1 of the DCASE 2024 challenge. The proposed model is based on CNN and is trained solely on the TAU Urban Acoustic Scene 2022 Mobile development dataset. Our model achieves accuracies of 46.02%, 49.54%, 53.89%, 56.43% and 59.08%, for 5%, 10%, 25%, 50%, and 100% subsets, respectively, with a MAC of 25.9M and memory usage of 88.2KB.

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] Florian Schmid, Paul Primus, Toni Heittola, Annamaria Mesaros, Irene Martín-Morató, Khaled Koutini, and Gerhard Widmer, "Data-efficient low-complexity acoustic scene classification in the DCASE 2024 challenge," 2024.

[2] T. Heittola, A. Mesaros, and T. 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.

Table I Class-Wise Accuracies of Splits (Baseline) in Percentage

Split	**Airport**	**Bus**	**Metro**	**Metro Station**	**Park**	**Public Square**	**Shopping Mall**	**Street Pedestrian**	**Street Traffic**	**Tram**	**Acc**
5	34.77	45.21	30.79	40.03	62.06	22.28	52.07	31.32	70.23	35.20	42.40
10	38.50	47.99	36.93	43.71	65.43	27.05	52.46	31.82	72.64	36.41	45.29
25	41.81	61.19	38.88	40.84	69.74	33.54	58.84	30.31	75.93	51.77	50.29
50	41.51	63.23	43.37	48.71	72.55	34.25	60.09	37.26	79.71	51.16	53.19
100	46.45	72.95	52.86	41.56	76.11	37.07	66.91	38.73	80.66	56.58	56.99

Table II Class-Wise Accuracies of Splits (Proposed) in Percentage

Split	**Airport**	**Bus**	**Metro**	**Metro Station**	**Park**	**Public Square**	**Shopping Mall**	**Street Pedestrian**	**Street Traffic**	**Tram**	**Acc**
5	42.13	59.9	32.56	43.91	68.38	15.05	57.61	31.65	68.89	40.17	46.02
10	42.47	68.11	37.78	40.74	66.87	24.68	71.55	28.82	74.14	40.24	49.54
25	52.4	65.15	52.15	47.71	75.82	35.35	51.78	41.72	67.88	48.92	53.89
50	51.08	65.28	45.92	46.73	77.03	39.79	69.15	36.46	72.22	60.67	56.43
100	53.38	69.66	51.07	52.92	78.38	40.16	56.29	47.00	79.39	62.06	59.08

[4] D. S. Park, et al., “SpecAugment: A simple data augmentation method for automatic speech recognition,” Interspeech, pp. 2613-2617, 2019.

[5] F. Schmid, T. Morocutti, S. Masoudian, K. Koutini, and G. Widmer, “Distilling the knowledge of transformers and CNNs with CP-mobile. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023),” 161–165. 2023.

[6] T. Morocutti, F. Schmid, K. Koutini, and G. Widmer, “Device robust acoustic scene classification via impulse response augmentation,” in 31st EUSIPCO, 2023.

[7] K. Y. Zhou, Y. X. Yang, and T. Xiang, “Domain generalization with mixstyle

[8] https://pytorch.org/docs/stable/quantization.html