SUBMISSION FOR DCASE 2024 TASK1: AN ASYMMETRIC RESIDUAL DEEP NEURAL NETWORK FOR LOW-COMPLEXITYACOUSTIC SCENE CLASSIFICATION

Technical Report

Chenhong Yan*, Yang Yu, Xiaohang Xiong

Northwestern Polytechnical Universi University
School of Marine Science and Technology,
Xi’an, 710072, China
chenhong.yan@mail.nwpu.edu.cn

ABSTRACT

This technical report describes the Sunshine Team24 ′ s submission for DCASE 2024 ASC Task 1, data-efficient low-complexity acoustic scene classification(ASC). Compared to Task 1 in the DCASE 2023 Challenge, the following aspects change in the 2024 edition: (a) Training sets of different sizes are provided. These train subsets contain approximately 5%, 10%, 25%, 50%, and 100% of the audio snippets in the train split provided in Task 1 of the DCASE 2023 challenge. A system must only be trained on the specified subset and the explicitly allowed external resources. (b)The model complexity is not part of the ranking system. The model's complexity is limited in terms of hard constraints. To this end, the memory requirement for model parameters is restricted to 128 kB, and the maximum number of multiply-accumulate operations (MACs) is restricted to 30 million. In this report, we present low-complexity systems for ASC that follow the rules intended for the task.

Index Terms—Low Complexity Network, Acoustic Scene Classification, Asymmetric convolution, depthwise separable convolution

1. INTRODUCTION

Acoustic scene classification tries to classify recordings in environments into a set of predefined classes. Deep neural networks have become a standard technique for this task[1]. However, the number of parameters required in state-of-the-art network models is usually more than a few million[1]. Hence, these solutions are very expensive to deploy on mobile phones or low-power-consumption devices. As a consequence, a low-complexity solution for acoustic scene classification is of great interest.

Deep networks have been applied successfully in vision, and the low-complexity solutions have been an active topic of research.As a recent example, Mobilenets[2][3] are deep learning networks that can reduce the number of parameters required while maintaining reasonable performance. Key features of these networks include depth-wise separable convolutions, and linear bottlenecks. Our solution for DCASE 2024 Task 1 builds on a lightweight multi-scale asymmetric residual convolutional neural networ which has high performance on the development dataset. Both depthwise separable convolution[3] and asymmetric convolution[5] are embodied in the structure. This depth-wise separable convolution has been widely used in lightweight networks and has been proven to be efficient[4]. The asymmetric convolution is adopted to extract time－frequency information individually and generate depth-level features. Thus, information complementarity can be achieved between different features, which can bring more useful information to the subsequent classification network and improve the classification performance of the network.

The contributions of this technical report are organized as follows: Chapter 2 describes the details of proposed algorithm framework. Chapter 3 describes the experiments and analysis results. Chapter 4 draws the conclusion.

2. PROPOSED METHOD

2.1. Datasets

The development dataset for DCASE 2024 challenge task1 is TAU Urban Acoustic Scenes 2022 Mobile development dataset[5]. The development set contains 230350 audio segments from 10 acoustic scenes. Duration of each segment is 1 second, in order to comply with the inference time and computational limitations imposed by the considered target devices.

This year, DCASE 2024 challenge task1 provides five pre-defined subsets/splits of the development-train dataset that are 100%, 50%, 25%, 10%, and 5% of the original development-train set's size. The 100% subset contains all segments (139,620) of the development-train split.

2.2 Preprocessing

The input features are extracted from the audio signals using a Short Time Fourier Transformation (STFT) with fft size of 4096 and the hop length of 500. We apply a Mel-scaled filter bank to end up with 256 frequency bins. We try two data augmentation methods in our system: Masking and MixStyle. MixStyle focuses on enhancing the diversity of feature representations by mixing style statistics from different instances.

2.3 Network architecture

Our proposed multi-scale asymmetric residual convolutional neural network(MAR-CNN) was based on the multi-scale asymmetric CNN with an attention mechanism(MA-CNN-A) model from[6]. Because of the low-complexity requirement, we first reduced the number of branches from the original model. Each branch employs convolution kernels of distinct sizes, i.e., 5, 9, and 17. These kernels ensure the extraction of multi-scale difference features and each branches further integrates asymmetric blocks. The group convolutions in parallel convolutional layers can be regarded as depthwise convolution, Both depthwise separable convolution and residuals[7] are embodied in the structure. Depthwise separable convolution and residuals in each branch is shown in Fig.3. The number of parameters that our models used is summarized in Table 1.

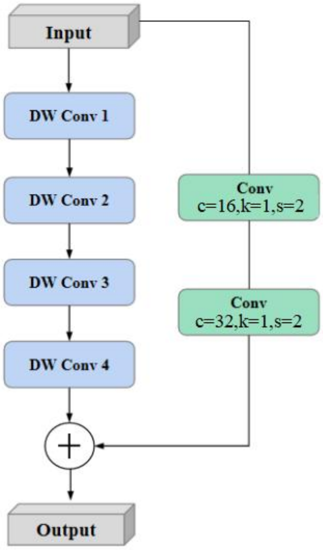

Figure 1: The structure of depthwise separable convolution and residuals. 'DW' representens depth-wise separable convolution.

Table 1: Summary model size for the proposed network

Model	MAR-CNN
Total params	31290
MAC	29675722

3. EXPERIMENTS

3.1. Experimental setup

We evaluated our proposed network on the development dataset. The sampling frequency is set to 32kHz, which is the same as baseline. Our batch size is 256. The input is 65 frames and 256 mel bins. We use Adam with a weight decay of 0.0001.

3.2. Results

Table 2 shows the acuracy with the development set for proposed networks. Table 3 shows the parameters and MMACs of our model and baseline.

Table 2 Accuracy of our model and Baseline

Subset	Baseline	Our model
5%	42.40%	45.71%
10%	45.29%	48.33%
25%	50.29%	52.07%
50%	53.19%	55.66%
100%	56.99%	59.80%
Overall	49.63%	52.31%

Table 3: Parameters and MMACs of our model and baseline

Model	Baseline	ours
Total params	61.148K	31.290K
MMACs	29.42M	29.68M

4. CONCLUSION

From the performance of the proposed small network, we can conclude that deep neural networks for acoustic scene classification can leverage depthwise separable convolutions and asymmetric convolutions to reduce the model size while maintaining reasonable performance.

5. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, ″Acoustic scene classification in dcase 2019 challenge: Closed and open set classification and data mismatch setups,″ 2019.

[2] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, ″Mobilenets: Efficient convolutional neural networks for mobile vision applications, ″ arXiv preprint arXiv:1704.04861, 2017.

[3] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, ″Mobilenetv2: Inverted residuals and linear bottlenecks,″ in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[4] Ding, Xiaohan, et al. "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[5] A. Mesaros, T. Heittola, and T. Virtanen, ″A multi-device dataset for urban acoustic scene classification, ″ in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop(DCASE2018), Nov 2018, p. 9–13.

[6] Yan, Chenhong, et al. "A Lightweight Network Based on Multi-Scale Asymmetric Convolutional Neural Networks with Attention Mechanism for Ship-Radiated Noise Classification." Journal of Marine Science and Engineering 12.1 (2024): 130.

[7] H. Hu, C.-H. H. Yang, X. Xia, X. Bai, X. Tang, Y. Wang,S. Niu, L. Chai, J. Li, H. Zhu, et al., ″A two-stage approach to device-robust acoustic scene classification,″ arXiv preprint arXiv:2011.01447, 2020.