DCASE 2021 TASK 1 SUBTASK A: LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION

Technical Report

Duc H. Phan *

University of Illinois at Urbana-Champaign
Illinois, USA
ducphan2@illinois.edu

Douglas L. Jones

University of Illinois at Urbana-Champaign
Illinois, USA
dl-jones@illinois.edu

ABSTRACT

Decomposing 2D convolution into time and frequency separable 1D convolutions produces a low-complexity neural network with good performance for acoustic scene classification. The final proposed network has roughly 41K parameters with a size of 75KB. It significantly outperforms the DCASE 2021 baseline network [1], with an accuracy of 64 percent on the development dataset [2].

***Index Terms*—** Low Complexity Network, Acoustic Scene Classification, depth-wise Separable Convolutions

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 [3]. However, the number of parameters required in state-of-the-art network models is usually more than a few million [3]. 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 [4, 5] 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 [5]. Our solution for DCASE 2021 Task 1A builds on the ResNet model from [6] which has high performance on the development dataset. Our proposed model is created by reducing the depth of the original model before replacing 2D convolution by depth-wise separable time and frequency convolutions. The rest of this report is organized as follows: First, a description of the development dataset is provided before introducing our proposed model. Next, the performance of the proposed method against the baseline is shown, followed by a conclusion.

2. DATA SET AND PREPROCESSING STEP

The DCASE 2021 Task 1 subtask A dataset contains recordings of 10 different acoustic scenes from 12 European cities with 4 recording devices [2]. From the original recording devices, 11 simulated devices are created by applying different impulse responses and dynamic compression ranges from recordings of Device A. The development dataset include three real Devices A, B, and C, and six simulated Devices S1-S6. In addition, the development dataset only include recordings from 10 cities. The acoustic scenes are grouped into 10 classes: airport, shopping mall, metro station, pedestrian street, public square, street traffic, tram, bus, metro, and park. 64 hours of 24-bit format recordings of single-channel audio at a sampling rate of 44.1kHz are provided in the development dataset.

In prepossessing steps, each recording was converted to log mel-band energy spectrograms with 128 mel bands. The number of samples in an analysis frame was 2048 with 50% hop interval. Each spectrogam was normalized into a range from 0 to 1 by its maximum and minimum values. Log-mel deltas and delta-deltas without padding were included as additional inputs into our models.

3. MODEL AND TRAINING

Our proposed model was based on the ResNet model from [6]. Because of the low-complexity requirement, we first reduced the depth and number of filters from the original model. After that, we decomposed the 2D convolutions into time and frequency separable convolutions: depth-wise separable convolution along the frequency axis, then another depth-wise separable convolution along the time axis preceding a 1x1 2D convolution. In our final model, the "compressed-Resnet model", the kernel size of our frequency filters is 7×1 and the time filters are 1×5. The number of parameters that our models used is summarized in Table 1

Our model was trained using Stochastic Gradient descent and warm restart similar the setting from [6] for 126 epochs. Mix-up augmentation [7] was employed during our training. After the training we selected the model with best accuracy on the validation dataset. After the training, the selected model was quantized into 16 bits floating points to further reduce the model size.

4. PERFORMANCE ON DEVELOPMENT DATA SET

There are two metrics for the task performance: accuracy, and multi-class cross-entropy. Accuracy will be calculated as macro-average (average of the class-wise accuracy for the acoustic scene classes). Multi-class cross-entropy (log loss) is used as a metric which is independent of the operating point [1]. This year the log loss metric is used for ranking, therefore we provide a post-

*This work utilizes resources supported by the National Science Foundation's Major Research Instrumentation program, Grant #1725729, as well as the University of Illinois at Urbana-Champaign.

Model	Compressed-ResNet
Total params	41,356
Trainable params	39,980
Non-trainable params	1,376
Non-zero params	36364
16 bit-float model size	75270 Bytes

Table 1: Summary model size for the proposed network

System	Outputs for normalization	Non-zero parameters	Total Size
task1a_1	top 2	36364	75270 B
task1a_2	top 3	36364	75270 B
task1a_3	top 2	36603	75756 B
task1a_4	top 3	36603	75756 B

Table 2: Summary of the submissions to DCASE 2021 Task 1A.

processing step at the output of the model to normalize the probability of each class to prevent the case where a model has a good accuracy while the value of log-loss is high.

Our post-processing steps go as follows. We select the top two or three outputs among the ten class-probability outputs of the models, then normalize the top two or three to 1.0 and replace the original values. We then normalize the 10 outputs such that they sum to 1. In this report we select the top 2 outputs for our normalization procedure.

The proposed system was trained and tested 10 times; the mean and standard deviation of the performance from these 10 independent trials are shown in the results table. The baseline model of DCASE 2020 task 1A is included for comparison [1]. As shown in Table 3, our proposed network outperforms the baseline system even though it is smaller in size. Our network also outperforms the baseline from DCASE 2020 Task A where the model size is not limited.

For DCASE 2021 Task 1A submission, we submitted four versions, two of them trained using the training and validation split provided by the development dataset. The others are trained using the entire development dataset. We also alternate between top 2 and top 3 outputs for our post-processing normalization in our submissions. Table 2 provide summary of the submissions.

System	Accuracy(%)	Log loss	Total Size
DCASE2021 Task 1A Baseline	47.7 ± 0.9	1.473 ± 0.05	90.3 KB
DCASE2020 Task 1A Baseline, Subtask A	54.1 ± 1.4	1.365 ± 0.003	19.4 MB
The Proposed network	63.54 ± 0.5	1.267 ± 0.009	73.5 KB

Table 3: Result of the proposed network in comparison with the systems provided by DCASE 2021 Task 1A.

5. CONCLUSION

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

6. REFERENCES

[1] I. Martín-Morató, T. Heittola, A. Mesaros, and T. Virtanen, "Low-complexity acoustic scene classification for multi-device audio: analysis of dcase 2021 challenge systems," *arXiv preprint arXiv:2105.13734*, 2021.

[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)*, 2020, submitted. [Online]. Available: https://arxiv.org/abs/2005.14623

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

[4] 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.

[5] 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.

[6] 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.

[7] H. Inoue, "Data augmentation by pairing samples for images classification," *arXiv preprint arXiv:1801.02929*, 2018.