DCASE 2020 TASK 1 SUBTASK B: 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

A deep network with depth-wise separable convolutions [1] and skip connections is introduced for low complexity acoustic scenes classification. The proposed network is not only more than 15 times smaller than the baseline convolution neural network [2] but also outperforms the baseline by two percents on average.

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 has become a standard techniques for this task [3]. However, number of parameters required in the state-of-the-art network models is usually more than few millions [3]. Hence, these solutions is very challenging 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 top of research. Recently 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 separable convolutions, depth-wise separable convolutions, and linear bottlenecks [5]. Motivated by Mobilenets, we introduce a network architecture that is 15 times smaller than the baseline system of DCASE 2020 task 1 subtask B while improving the performance approximately by two percents. Our network applied the key insights from Mobilenetv2 [5]. The rest of this report is organized as follows: First, description of the development data set is provided before introducing our proposed model. Next, performance of the proposed method against the baseline is shown, followed by conclusion.

2. DATA SET AND PREPROCESSING STEP

The DCASE 2020 Task 1 subtask B dataset contains recordings of 10 different acoustic scenes from 12 European cities [2]. The acoustic scenes are grouped into three classes: indoor, outdoor, and transportation. All recordings are binaural, 48kHz 24-bit format, and from a single recording device. The development data set has 40 hours of recording from 10 different cities. 70 percents of the development data set is used for training, while 30 percent is withhold as a test set. Recordings from the same location can only in the training set or test set but not both.

In prepossessing steps, first channel of a recording is converted to log mel-band energies spectrogram with 40 mel bands. The number of samples in a analysis frame is 2048 (40 ms) with 50% hop interval. The data set is normalized frequency-wise across time. The mean and standard deviation are estimated from the training set.

3. MODEL AND TRAINING

The log mel-band energies spectrogram feature, $X \in \mathbf{R}^{F\times T}$, is given as the input to the proposed network, where F is 40 mel bands and T is 498 analysis frames (10 s audio recording). The proposed network structure is presented in Table 1. Layer names are Keras definitions [6]. Note that after each batch normalization (BN) layer, a rectifired linear unit is attached. The configurations of each the layer are provided in Table 2. Using the model size calculation provided by the task, our model has 6979 parameters in total with the size of 27916 bytes or nearly 27.26 KB, while number of nonzero parameters is 6944 with the size of 27776 bytes or nearly 27.13 KB.

The network was trained for 200 epochs with a batch size of 64. ADAM optimizer [7] was used with a learning rate of 0.0001 and a dropout rate of 0.05. The training model that has smallest classification error over the validation set is selected. We submit two outputs this model and another two outputs from the extension of this model where the number of filters in convolution layers were doubled and a dropout rate is set to 0.2. The next section present the performance of the proposed method on the development dataset.

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 [2].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 table3. The baseline model and subtask A baseline system and

[*] 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.

Layer	Connected to
$Input$	
$Conv2D_1$	$Input$
$DepthwiseConv2D_1$	$Conv2D_1$
$Concatenate_1$	$DepthwiseConv2D_1$ $Input$
$BatchNormalization_1$	$Concatenate_1$
$AveragePooling2D_1$	$BatchNormalization_1$
$Conv2D_2$	$AveragePooling2D_1$
$BatchNormalization_2$	$Conv2D_2$
$Dropout_1$	$BatchNormalization_2$
$DepthwiseConv2D_2$	$Dropout_1$
$DepthwiseConv2D_3$	$DepthwiseConv2D_2$
$Concatenate_2$	$DepthwiseConv2D_2$ $DepthwiseConv2D_3$
$BatchNormalization_3$	$Concatenate_2$
$Conv2D_3$	$BatchNormalization_3$
$BatchNormalization_4$	$Conv2D_3$
$GlobalMaxPooling2D_1$	$BatchNormalization_4$
$Dropout_2$	$GlobalMaxPooling2D_1$
$Dense_1$	$Dropout_2$
$Output$	$Dense_1$

Table 1: Network connection of the proposed model. Layer names are Keras layer definitions

Layer	Configuration
$Conv2D_1$	filters=32, kernel_size=(4, 1)
$DepthwiseConv2D_1$	kernel_size=(1, 5)
$Conv2D_2$	filters=32, kernel_size=(1, 1)
$DepthwiseConv2D_2$	kernel_size=(5,1)
$DepthwiseConv2D_3$	kernel_size=(1, 5)
$Conv2D_3$	filters=32, kernel_size=(1, 1)
$Dense_1$	units=3

Table 2: Configurations of layers if necessary.

a minified version of it are included for comparision [2]. Clearly, our proposed network outperforms the baseline system even though it is much smaller in size. Our network is very close to the performance of the subtask A baseline which used a huge pre-trained OpenL3[8] network to calculate audio embedding.

5. CONCLUSION

From the performance of the proposed small network, we can conclude that deep neural network for acoustic scene classification can leverage depth-wise separable layer convolutions and bottleneck layers from vision in order to improve performance while keeping small model size.

6. REFERENCES

[1] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[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] A. Gulli and S. Pal, *Deep learning with Keras*. Packt Publishing Ltd, 2017.

[7] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[8] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen and learn more: Design choices for deep audio embeddings," in *IEEE Int.˜Conf.˜on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 3852–3856. [Online]. Available: https://ieeexplore.ieee.org/document/8682475

System	Accuracy	Log loss	Audio Embedding	Total Size
DCASE2020 Task 1 Baseline, Subtask A	89.8 ± 0.3 %	0.266 ± 0.3	17.87 MB	19.12 MB
Modified DCASE2020 Task 1 Baseline, Subtask A	88.9 % ± 0.3 %	0.298 ± 0.003	840.6 KB	985.8 KB
DCASE2020 Task 1 Baseline, Subtask B	87.3 ± 0.7	0.437 ± 0.045	0 B	450 KB
The Proposed network, Subtask B	89.5±0.2	0.287 ± 0.006	0 B	27.26 KB

Table 3: Result of the proposed network in comparison with the systems provided by Dcase2020 task 1.