COMP.SGN.240 Advanced Signal Processing Laboratory DCASE 2020 Task 1: Subtask B Report

Fareeda Mohammad, Nardos Estifanos

November 2023

1 Introduction

The DCASE 2020 Task 1 challenge is focused on categorizing acoustic scenes to appropriate environment recorded using 4 devices with 11 more simulated devices. The task has 2 subtasks; Acoustic scene classification with 10 environments, and low complexity classification with generalized 3 places (indoor, outdoor, and transportation). The provided baseline uses CNN with 10-second log-mel band energies, and achieves 87.3% accuracy [12] with 450KB.

2 Literature Review

2.1 Overview of Low-Complexity Models for ASC

Acoustic scene classification is a task of categorizing sounds produced by or in some well defined and distinct physical source [1, 9]. The aim of this task is to allow machines listen and understand surrounding sounds, which is crucial in responding appropriately and can be applicable in robotic systems[4], environmental monitoring, smart systems (Home, City, Building)[14], and more. State-of-the-art performance has been achieved following deep learning usage in the field, in particular, Convolutional Neural Networks proved to be effective[9]. However, edge devices and low-powered (memory, and CPU computational power) devices struggle to run modern day models, hence the need for low-complexity classifiers[10]. This means the target is no longer about accuracy rather size of network as well.

To realize these classifiers, there are several proven approaches, but in general, they can be categorized into three based on what they focus on. Input focused, model/network focused, and model compression focused.

Input-focused approaches try to achieve optimized learning by using optimal feature representations. Some of the methods argue that learning multiple features from raw signal as opposed to common transforms like Log-mel or STFT promote peak performance which helps train simple models with distinct input[5]. In addition, existing works utilize multiple timescale features, however interleaving wavelet scattering based simple mixing helps achieve low-complexity.[8]

More works focus, however, on the network itself. Most works [9, 10, 11], utilize some form of residual connections in their design, since they help learn complex features with fewer parameters[6]. Dilated convolutions, depth-wise separable CNNs, and mobile blocks result in reduction of network parameter, together with post-training quantization [9] achieved up to 81 KB.

Another rising approach is post-training model compression by quantization [9], pruning, knowledge distillation to list a few. Prunning selectes important connections which significantly reduce network size, and this method is popular choice [12]. Knowledge distillation which works by having complex model as a teacher, and a low-complexity model as a student, and knowledge is trickled down to the student using soft labels[13] . Using multi-representation along with knowledge distillation can result in higher accuracy while keeping complexity low[5]

To summarize, the rise in demand to use acoustic models in edge devices requires low-complexity models and combining different strategies helps achieve small models with out compromising performance. *In this work, we try to utilize depthwise separable CNNs [3] to achieve low-complexity high accuracy model.*

3 Dataset

From the task webpage, the dataset for subtask B is TAU Urban Acoustic Scenes 2020 3Class and development set[7] which is around 40GB is recommended to use. However, due to hardware limitation, and the time it takes to experiment our models, only 1000 audio (.wav) files were randomly selected from the 3 classes: indoor, outdoor, and transportation. In total it would be around 8.1GB. This was split into 3 sets: train(75%), validation(10%), and test(15%). The split has fair distribution as it can be seen in Figure 1.

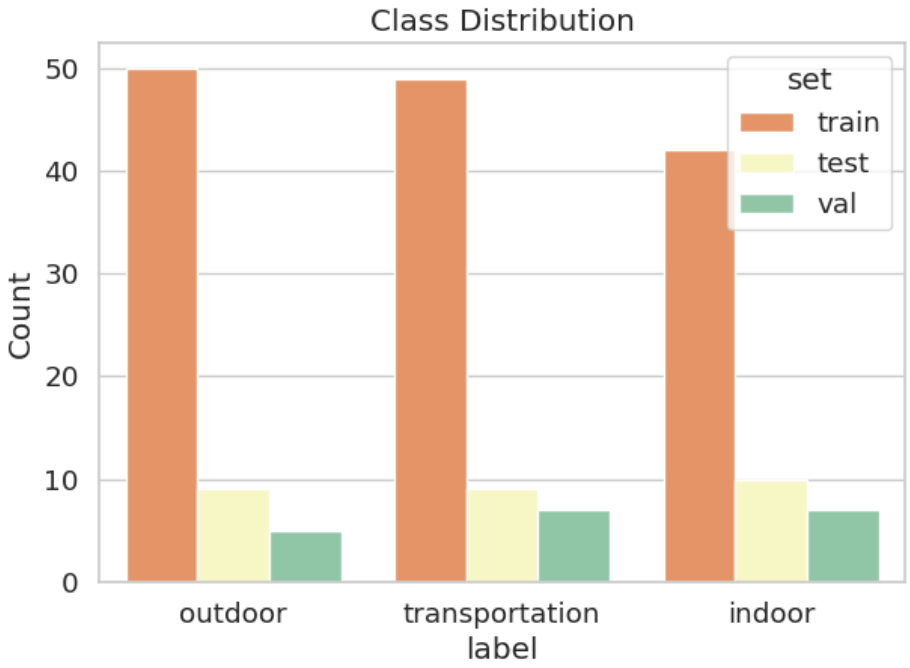


Figure 1: Dataset class distribution per split set.

4 Methodology

4.1 Experiment and Results

Since the provided baseline script was not fully working as expected in our setup, after downloading the dataset, custom shell scripts were used to extract all the zip files, and store them into one folder. Then the 10 acoustic scenes were grouped into 3 classes following the instruction, i.e indoor contains airport, metro station and so on. Similarly, for the baseline implementation, specification from the instruction is followed and built, trained, **evaluated** in simplified manner, i.e with out k-fold validation. However, our model also follows similar pipeline.

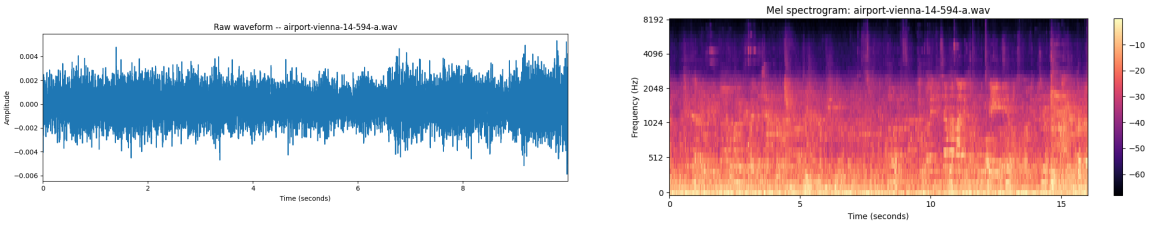
To train the models, Narvi Cluster of GPUs were used.

4.1.1 Feature Extraction

The subtask B uses 40 log mel-band energy features extracted with 40ms window and 50% hop size. Librosa is used to extract and visualize this feature, see sample in Figure 2.

4.1.2 Baseline Model

The baseline model uses CNN based approach with the following specification, which is shown in Figure 3. Input is 40x501 shape, CNN1 (32, (7,7)), Maxpool ((5,5)), CNN2 (64, (7, 7)), Maxpool ((4, 100)). Following the CNN layers batch normalization and ReLu layers exist, and after the Maxpool layers a



(a) Sample Raw audio

(b) Sample log mel-band energy feature

Figure 2: Sample input feature

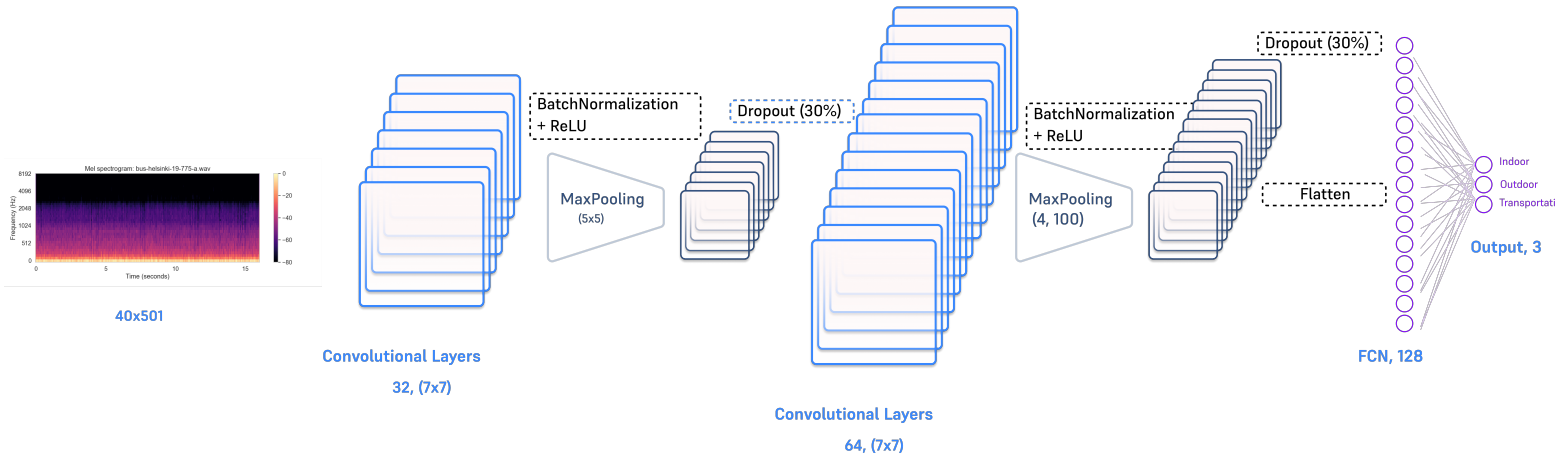
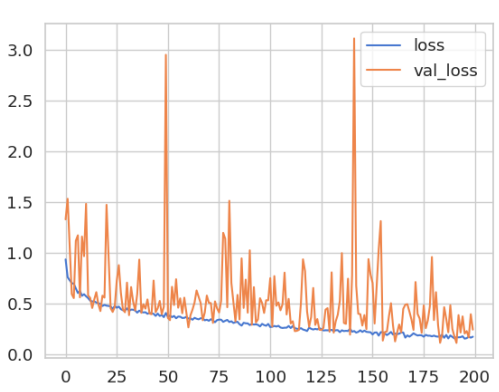


Figure 3: Baseline model architecture

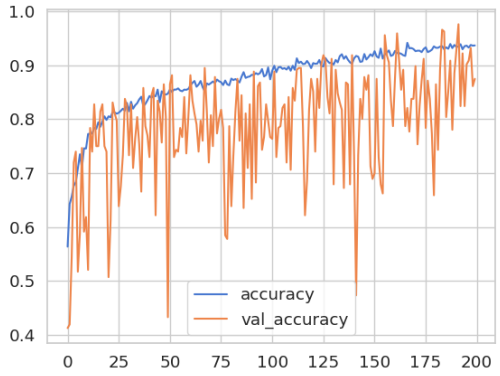
dropout of 30%. Flattened output is fed to 100 dense neurons with similar dropout. The output layer uses softmax activation for the 3 classes.

Model is compiled with categorical cross-entropy loss and Adam optimizer (lr=0.001), and trained with batch of 16, data shuffling in between 200 epochs.

In our setup and training of the baseline model using 3000 wav files, a loss of 0.15, with accuracy of 95.3% was obtained. And the model size reported by using similar script provided in the baseline repository, was 450.82 KB, and non-zero trainable parameters were 115,411. Training history is shown in Figure 4



(a) Training loss (x-axis is epoch, y-axis is loss)



(b) Training Accuracy (x-axis is epoch, y-axis is accuracy)

Figure 4: Baseline model training history

4.1.3 Our Implementation

Inspired by related works, using separable CNN layers reduced the model complexity. The same input feature is used, and the architecture is shown below in Figure 5. To keep model size below 500KB, narrow architecture was used and small filters. With this network only 40 epochs are needed to reach convergence and training is significantly faster which resulted in higher accuracy than the baseline.

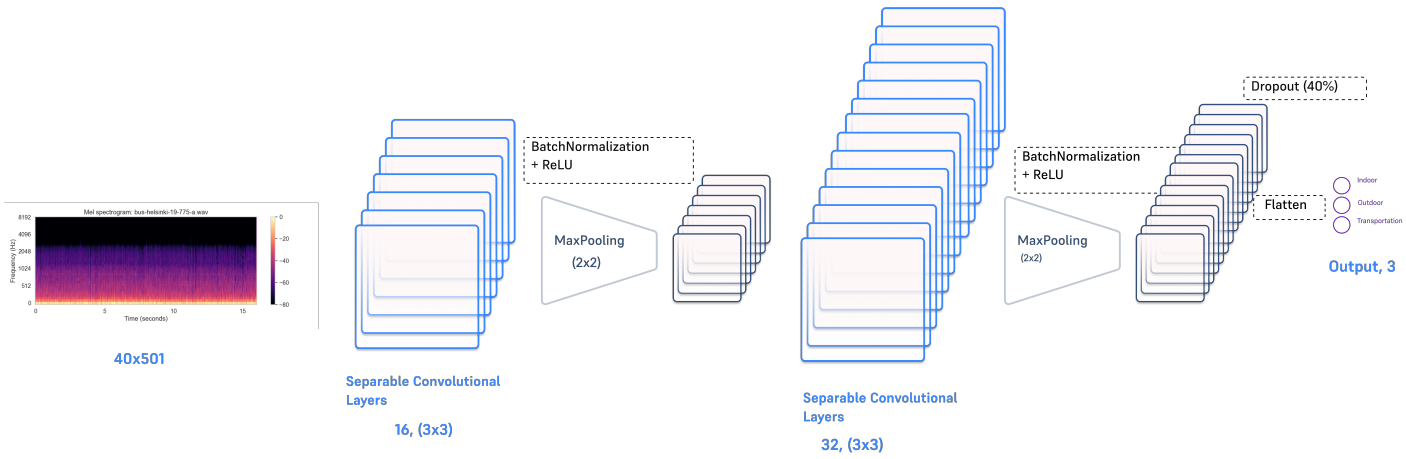
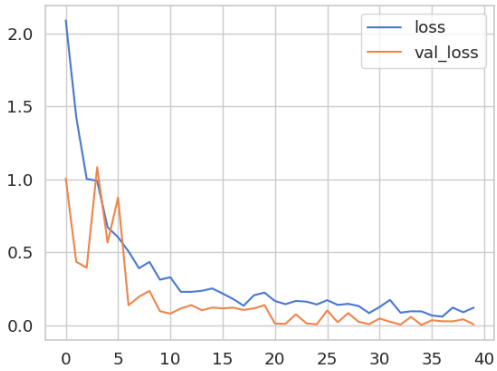


Figure 5: Our Implementation Architecture

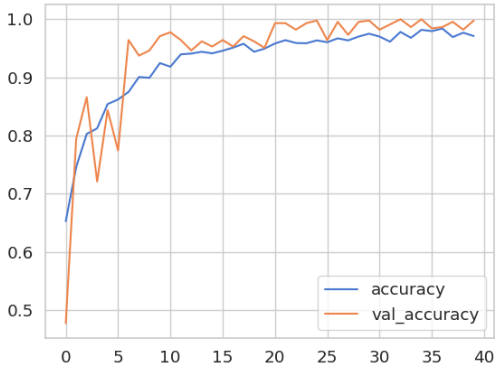
Our model was built using tensorflow's Separable2D(...) function, and after the flatten Dense layer is removed unlike the baseline model, which dropped the complexity under the required amount. Then, it was compiled with Nadam optimizer[2] which leads to improved convergence. As it can been observed in 6a , training is more stable compared to baseline and converges well. **Loss of 0.0033 and accuracy of 99.55% with 371.86 KB and 95196 non-zero parameters was achieved**. However, since the full dataset was not utilized, this does not indicate that our implementation exceeds the challenge's leader boards.

4.1.4 Performance on test set

From the extracted audio files, 3 wav files from each classes were collected in to a test folder. Using the trained model with these potentially unseen audio files, we get the following results.(See Figure 7)



(a) Training loss (x-axis is epoch, y-axis is loss)



(b) Training Accuracy (x-axis is epoch, y-axis is accuracy)

Figure 6: Our model training history

5 Summary

5.1 Subjective Analysis on Baseline

The baseline model utilizes normal CNN layers, with the filter numbers used, and large kernel size (7x7) which captures high level features, but it make it computationally expensive. In our model, dense layer after flatten is removed and it significantly gave rise to low complexity model, hence, baseline model might also be affected by this layer.

In our model as well as baseline, transportation class gets mistaken with outdoor. It's understandable even to our ears it might not feel distinctive enough.

To summarize, utilizing depthwise separable CNN layers reduces complexity, improves speed inference of models. Avoiding fully connected neural networks encourages in similar manner.

References

[1] Emmanouil Benetos, Dan Stowell, and Mark D. Plumbley. Approaches to complex sound scene analysis. 2018.

[2] Kushal Chakrabarti and Nikhil Chopra. A state-space perspective on the expedited gradient methods: Nadam, radam, and rescaled gradient flow. *2022 Eighth Indian Control Conference (ICC)*, pages 31–36, 2022.

[3] François Chollet. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2016.

[4] Brian P. Clarkson, Nitin Sawhney, and Alex Paul Pentland. Auditory context awareness via wearable computing. *Energy*, 1998.

[5] Liang Gao, Kele Xu, Huaimin Wang, and Yuxing Peng. Multi-representation knowledge distillation for audio classification. *Multimedia Tools and Applications*, 81:5089 – 5112, 2020.

[6] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[7] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. Tau urban acoustic scenes 2020 mobile, development dataset. 2020.

[8] Xing Yong Kek, Cheng Siong Chin, and Ye Li. An intelligent low-complexity computing interleaving wavelet scattering based mobile shuffling network for acoustic scene classification. *IEEE Access*, 10:82185–82201, 2022.

[9] Aswathy Madhu and K. Suresh. Atresnet: Residual atrous cnn with multi-scale feature representation for low complexity acoustic scene classification. *Circuits, Systems, and Signal Processing*, 41:7035 – 7056, 2022.

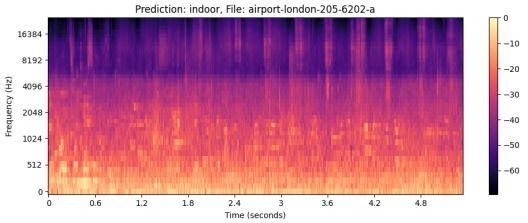
[10] Lam Dang Pham, Dusan Salovic, Anahid N. Jalali, Alexander Schindler, Khoa Tran, Hai Canh Vu, and Phu X. Nguyen. Robust, general, and low complexity acoustic scene classification systems and an effective visualization for presenting a sound scene context. *ArXiv*, abs/2210.08610, 2022.

[11] Soonshin Seo, Junseok Oh, Eunsoo Cho, Hosung Park, Gyujin Kim, and Ji-Hwan Kim. Tp-mobnet: A two-pass mobile network for low-complexity classification of acoustic scene. *Computers, Materials & Continua*, 2022.

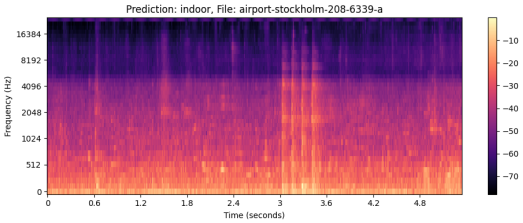
[12] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, October 2015.

[13] Jee weon Jung, Hee-Soo Heo, Hye jin Shim, and Ha jin Yu. Knowledge distillation in acoustic scene classification. *IEEE Access*, 8:166870–166879, 2020.

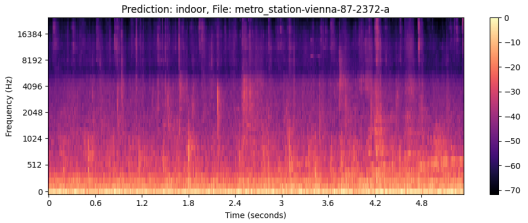
[14] Chengyun Zhang, Haisong Zhan, Zezhou Hao, and Xinghui Gao. Classification of complicated urban forest acoustic scenes with deep learning models. *Forests*, 2023.



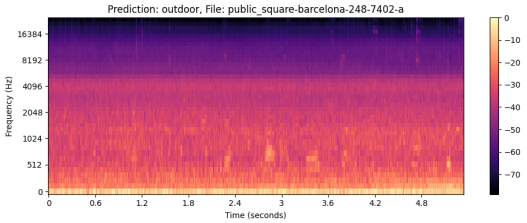
(a) Originally Indoor



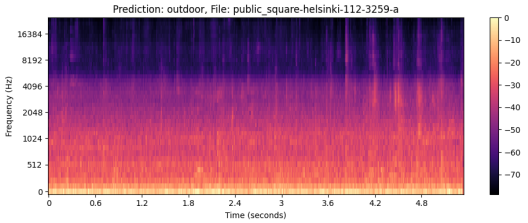
(b) Originally Indoor



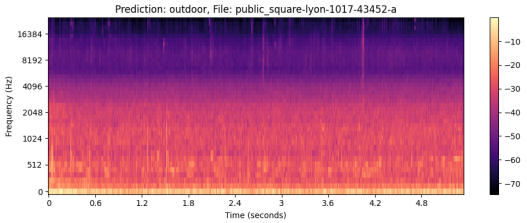
(c) Originally Indoor



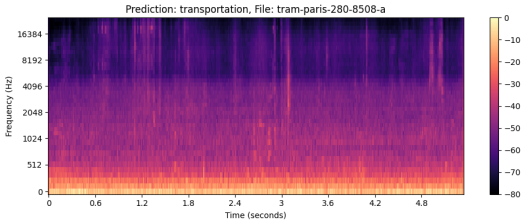
(d) Originally Outdoor



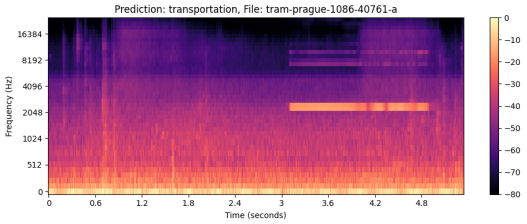
(e) Originally Outdoor



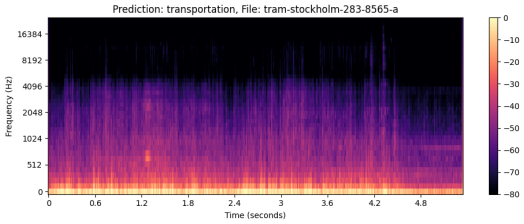
(f) Originally Outdoor



(g) Originally transportation



(h) Originally transportation



(i) Originally transportation

Figure 7: Our model performance on unseen test wav files 3 from each class.