Comparing light-weight CNN model for acoustic scene classification

Soyoung Lim, Il-Youp Kwak
Department of Statistics, Chung-Ang University, Seoul, Republic of Korea

Introduction

In recent years, acoustic scene classification (ASC) has attracted widespread attention in the Audio and Acoustic Signal Processing (AASP) community. ASC aims to classify a test recording sound into predefined classes that characterizes the environment in which it was recorded. The IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) takes place every year. Our task aims to classify audio into three classes based on low-complexity solutions.

- Our proposed system ranked 7th in the competition for DCASE 2020 task 1B.

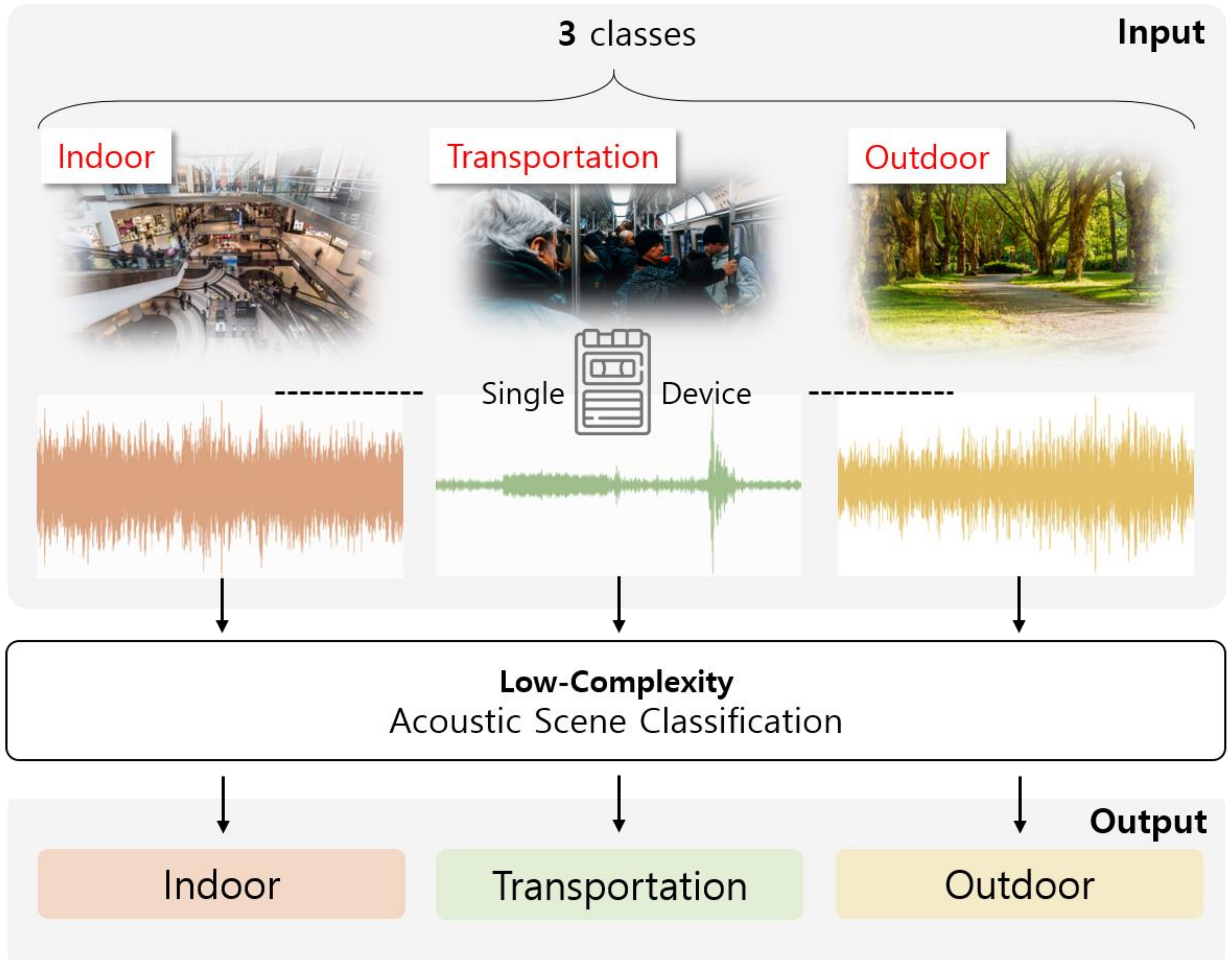

Figure 1. Overview of the ASC system

- We use Log Mel-spectrogram, Deltas-Deltadeltas features in ASC systems.
- we reduced the number of ResNet layers.
- the model size was reduced by using the depthwise separable convolution used in MobileNet v1, bottleneck inverted residual block used in MobileNet v2, and Quantization.

Methods & Materials

Data

DCASE 2020 Task1 B(TAU Urban Acoustic Scenes 2020 3Class) Development Dataset

System Architecture

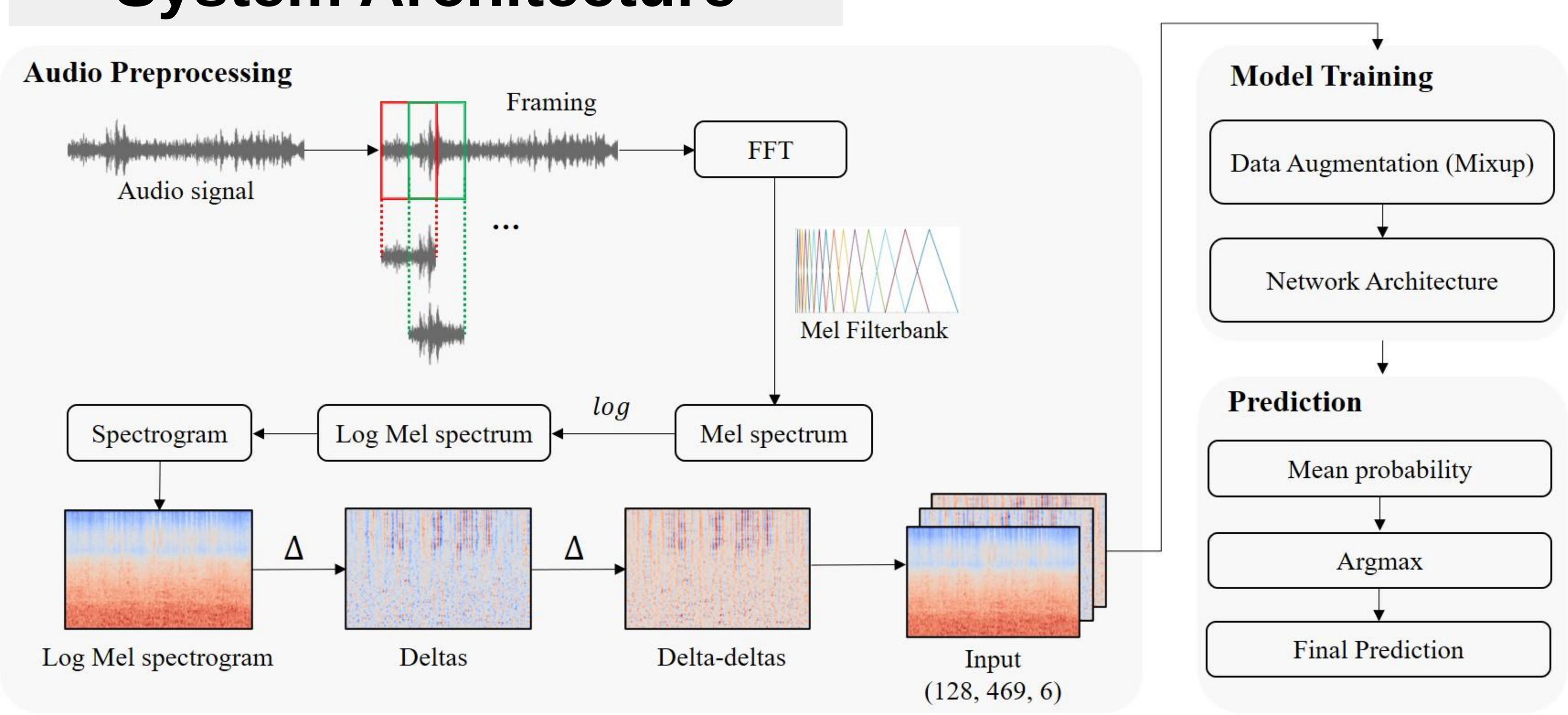

Figure 2. Proposed system architecture

Audio Preprocessing

Log Mel-spectrogram, Deltas-Deltadeltas

Audio preprocessing Parameter	
Sampling rate	48,000 Hz
audio channel	binaural (2)
n_fft	2,048
hop_length	1,024
n_mels	128

Table 1. Audio preprocessing parameter

Model Train

Model training parameter	
Data Augmentation	Mixup
Loss	Categorical crossentropy
Optimizer	Adam
Evaluation metric	Categorical Accuracy
Learning rate scheduler	Sigmoidal decay function
Batch size	64
Epoch	100

Table 2. Model training parameter

Model Architecture

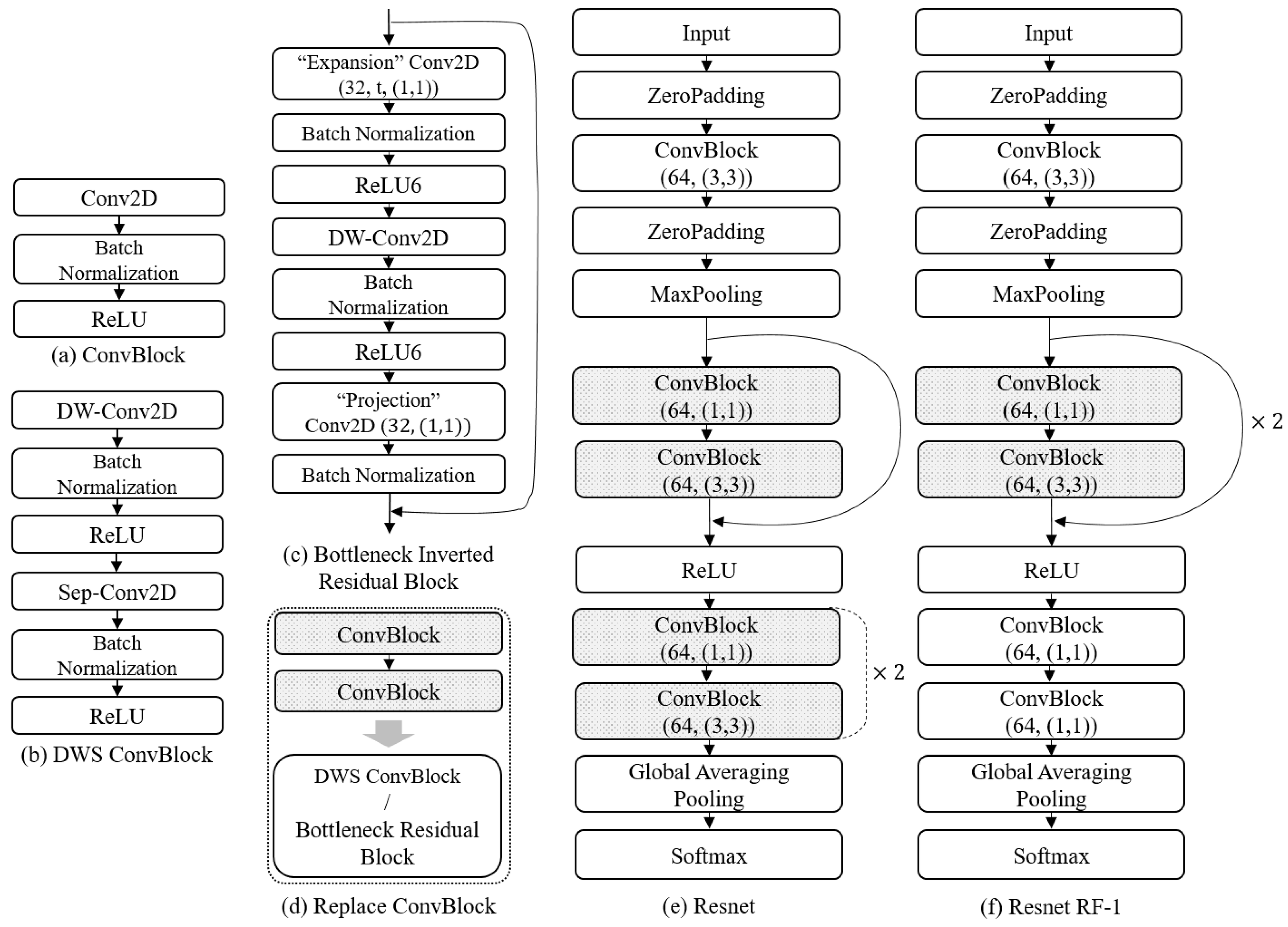

Figure 3. Model Architecture

Results

- **"ResNet":** the ResNet architecture from Figure 3 (e)
- **"ResNet RF1"** : the ResNet architecture with 1x1 last conv block from Figure 3 (f)
- **"DWS"**: the architecture with a DWS convolution layer from Figure 3 (b)
- **"BIR"**: the architecture with a Bottleneck Inverted Residual layer convolution layer from Figure 3 (c)
- **"Q"**: 16bit Quantization

	Model	Test accuracy (%)	Model Size (KB)
1	ResNet	95.05	503
2	DWS-ResNet	94.17	78.51
3	BIR-ResNet	**94.67**	**95.64**
4	ResNet RF-1	95.15	375
5	DWS-ResNet RF-1	94.15	92.01
6	BIR-ResNet RF-1	93.88	97.01
7	Q-ResNet	95.05	255
8	Q-DWS-ResNet	94.17	42.76
9	Q-BIR-ResNet	**94.64**	**52.07**
10	Q-ResNet RF-1	95.13	191
11	Q-DWS-ResNet RF-1	94.12	49.51
12	Q-BIR-ResNet RF-1	93.81	52.51

Table 2. Results

Conclusions

- The model considering low-complexity was similar or slightly inferior to the performance of the base model, but the model size was reduced from 503 to 42.76 KB.
- Therefore, we confirmed that the DWS convolution and BIR convolution are effective in reducing the model size while maintaining the performance of the model.

This work was supported by the National Research Foundation of Korea (NFR) grant funded by Ministry of Science and ICT (2020R1C1C1A01013020)