A Residual CNN with RepConv2d and LearnablePooling for Acoustic Scene Classification

Technical Report

Mohamad Mahdee Ramezanee

Sharif University of Technology
Electrical Engineering Department
Azadi Avenue
Tehran, 1458889694, Iran
ramezani.mm@ee.sharif.edu

Hossein Sharify

Sharif University of Technology
Electrical Engineering Department
Azadi Avenue
Tehran, 1458889694, Iran
sharifir.hossein@ee.sharif.edu

Amir Mohammad Mehrani Kia

Sharif University of Technology
Electrical Engineering Department
Azadi Avenue
Tehran, 1458889694, Iran
mehranikia.amohammad@ee.sharif.edu

Behnam Raoufi

Sharif University of Technology
Electrical Engineering Department
Azadi Avenue
Tehran, 1458889694, Iran
behnam.raoufi93@sharif.edu

ABSTRACT

The objective of the acoustic scene classification task is to categorize audio recordings into one of ten predetermined environmental sound categories, such as urban parks or metro stations. This report to Task 1 of the DCASE 2025 Challenge, which emphasizes developing data-efficient, low-complexity systems for acoustic scene classification, addressing real-world constraints like limited training data and device mismatches [1]. Our model is designed with a reparameterizable convolutional structure that unifies multiple asymmetric kernels into a single efficient layer during inference, enabling both rich spatial representation and computational efficiency. It further integrates a novel attention-guided pooling strategy and a hybrid normalization scheme to enhance feature discrimination and stability throughout the network. Finally, we utilized ensemble learning of the newly defined teacher models and minimized the KL divergence between the student and teacher models to improve the results.

***Index Terms*—** efficient layer,ensemble teachers, acoustic scene classification, low-complexity

1. INTRODUCTION

The DCASE 2025 Challenge Task: Low-Complexity Acoustic Scene Classification with Device Information focuses on classifying audio recordings into one of ten predefined acoustic scene classes while addressing several key challenges. These include managing recording device mismatches, adhering to strict low-complexity constraints, achieving data efficiency with limited training data, and developing models that can utilize device-specific information to improve performance while maintaining generalization to unseen devices.

Main Challenges

1. Recording Device Mismatches

The task requires systems to handle variations across different recording devices, including those not present in the training set, such as devices D, S7–S10 in the evaluation set.

2. Low-Complexity Constraints

Models must be designed with a maximum of 128 kB of parameters (equivalent to 128K parameters for int8, 64K for float16, 32K for float32) and a maximum of 30 million MACs per 1-second inference, reflecting the computational limitations of target devices like Cortex-M4

3. Data Efficiency

Participants are limited to using only 25% of the development set for training, totaling 18 hours of data, which encourages the adoption of pre-training and other data-efficient learning strategies

4. Device-Specific Models

The task allows for the use of device ID information at inference time, enabling the development of per-device fine-tuned models. However, systems must also demonstrate generalization to unknown devices, balancing specialization with broad applicability.

2. APPROACHES

The steps taken to train the final model and generate the results can be summarized in three main stages, which will be explained in the following sections.

2.1. IR Augmentation

Table 1: Mapping of RMS Energy Levels to Number of Audio Augmentations designed based on the available computation and memory for data augmentation.

Rms Energy Range	0.00 ≤ Energy ≤ 0.02	0.02 < Energy ≤ 0.05	0.05 < Energy ≤ 0.10	0.10 < Energy ≤ 0.15	0.15 < Energy ≤ 0.20	0.20 < Energy ≤ 0.25	0.25 < Energy ≤ 0.30	Energy > 0.30
#Augmentations	0	5	7	9	11	13	16	20

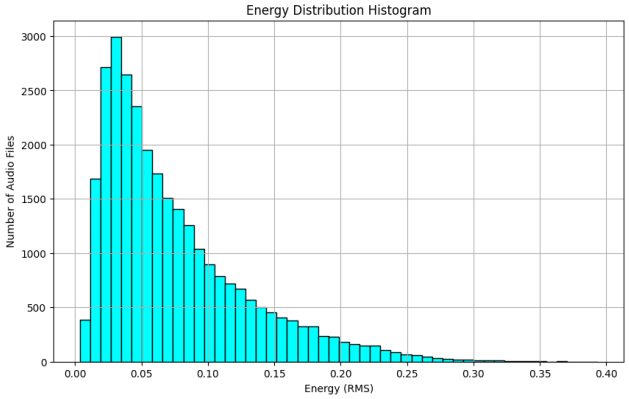

Figure 1 : Histogram of Audio File Energy Levels Based on RMS Values

The audio augmentation process involves loading WAV files from a dataset [2], converting them to mono, and resampling to 48 kHz if necessary. Based on the RMS energy of each audio file, a dynamic number of augmentations is determined Based on the RMS energy of each audio file, a dynamic number of augmentations is determined (see Table 1 and figure 1). Selected impulse responses(IRs) from a directory [3] are convolved with the audio to generate augmented versions, with RMS normalized to match the input. Augmented files are saved with unique names, and their metadata, including filename, label, and device, is recorded in a CSV file for tracking.

2.2. Model Design

Our student network is a convolutional neural network designed for image classification, incorporating RepConv2d, ResidualNormalization, and LearnablePooling modules. It begins with input normalization and two RepConv2d layers for initial feature extraction with downsampling. Three residual stages, each with DSFlexiNetBlocks, apply expansion, spatial convolution, and skip connections to enhance feature learning. The LearnablePooling module combines attention-based and global average pooling for robust feature aggregation. Finally, a classification head with normalization, dropout, and a linear layer outputs class predictions. The device-specific performance is detailed in Table 2.

Table 2 : Device-Specific performance metric for the trained student model in all samples

Device	Accuracy(%)
a	59.6
b	55
c	57.3
s1	53
s2	54.2
s3	57.5
others	53.4

2.3 Design Teachers

Table 3: Performance metrics for the trained teachers and their combination.

Name Of Models	Accuracy(%)	Log Loss
DSFlexi_ver_a	54.9	1.31
DSFlexi_ver_m1	54.0	1.27
DSFlexi_ver_m2	55.7	1.21
CP-mobile [4]	54.2	1.47
CP-ResNet [4]	52.0	1.35
Ensemble models	63.0%	1.02

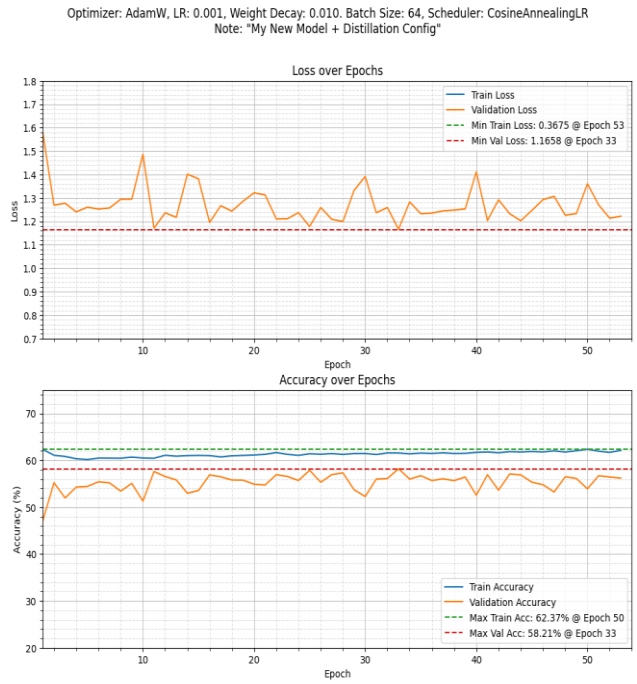

Figure 2: The accuracy and loss curves of the globally trained student model across all devices are shown. This pre-trained model is used for fine-tuning the student models for each individual device.

The knowledge distillation process, depicted in Figure 2, uses log-mel spectrogram features as input for multiple pretrained CNN-based teacher models to generate teacher logits. These logits guide the training of a student CNN model, which takes the same input and produces its own predictions. The training optimizes a combined loss function, balancing classification loss (cross-entropy with ground truth) and knowledge distillation loss (KL divergence between softened teacher and student outputs, using different temperature factors). This approach achieving a final accuracy of 58.2% for the student CNN model trained on all devices.

3. SUBMISSIONS

Finally, by modifying the hyperparameters of the final trained network and adjusting the coefficients of the distillation loss, we created three submissions. The results and all network codes are provided in the YAML files and the GitHub link [5].

4. REFERENCES

[1] https://dcase.community/challenge2025/task-low-complexity-acoustic-scene-classification-with-device-information
[2] https://zenodo.org/records/6337421
[3] https://www.kaggle.com/datasets/mahdyr/micirp
[4] Khaled Koutini, Hamid Eghbal-zadeh, and Gerhard Widmer, "Receptive f ield regularization techniques for audio classification and tagging with deep convolutional neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1987–2000, 2021.
[5] https://github.com/Mah-De/DCASE2025-Task1