# Mini-SegNet and Low-Complexity MobileNet for Acoustic Scene Classification

Technical Report

Ge-Ge Bing[1], Yun-Fei Shao[1], Zhi Zhang[2], Wei-Qiang Zhang[1]
1. SATLab, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
2. School of Computer Science, Beijing Institute of Techonology, Beijing 100081, China
bgg19@mails.tsinghua.edu.cn, shaoyf@tsinghua.edu.cn, 1120202775@bit.edu.cn, wqzhang@tsinghua.edu.cn

**ABSTRACT**

This report details the architecture we used to address task 1 of the DCASE2023 challenge. The goal of the task is to design an audio scene classification system for device-imbalanced datasets under the constraints of model complexity. Our architecture is based on (1) SegNet, applying structured pruning and quantization to reduce model complexity; (2) MobileNet with an additional frequency split block. Log-mel spectrograms, delta, and delta-delta features are extracted to train the acoustic scene classification model. Mixup, random crop, time and frequency domain masking are used for data augmentation. The proposed system achieves higher classification accuracies and lower log loss than the baseline system. After model compression, our single MobileNet model achieves an average accuracy of 51.3% with only 7.946K parameters, and 3.972M Multiply–Accumulate Operations (MACs), while pruned SegNet gets to an average accuracy of 54.46% with 46. 232K parameters and 19.466M MACs.

***Index Terms***— Acoustic scene classification, SegNet, MobileNet, data augmentation, model compression

## 1. INTRODUCTION

The aim of Task 1 in the DCASE2023 challenge is to recognize ten acoustic scene classes using a low-complexity classification model [1]. Acoustic Scene Classification (ASC) involves identifying the type of environment or scene from a given audio. The development dataset for Task 1 provides audio recordings from ten different acoustic scenes in ten European cities, such as airports, shopping malls, and metro stations. The task requires accurately assigning labels to the corresponding scenes while limiting the model's complexity to adapt to different devices. Thus, it is necessary to balance the reduction of model parameters and the potential decrease in model accuracy. Additionally, approaches such as spectrum correction, aggressive regularization, and augmentation should be employed to deal with acoustic scenes recorded and simulated with multiple devices.

In recent years, convolutional neural network (CNN) based models [2-4] have achieved impressive performance in ASC. However, these models often have a high complexity and a large number of parameters. To achieve a balance between low complexity and high accuracy, we implemented MobileNet [5] in Task 1. MobileNet uses depthwise separable convolutions to decompose standard convolutions into depthwise and pointwise convolutions, significantly reducing the number of parameters and computation required. When compared to GoogleNet and VGG, MobileNet performs similarly in terms of accuracy, but has reduced the number of parameters by two orders of magnitude compared to VGG 16, resulting in a significant reduction in complexity. Our model is based on the latest version of MobileNet, MobileNetV3 [6], which is both more accurate and efficient. Additionally, we optimized our mini-SegNet [7-9] in 2022 by applying model pruning, further reducing the complexity of the mini-SegNet model.

The remainder of this report is structured as follows. Section 2 provides an overview of our model from three different perspectives: data pre-processing, the proposed model, and data augmentation. In Section 3, we present the results of extensive experiments to validate the accuracy and performance of our model. Finally, we present our conclusions in Section 5.

## 2. THE SYSTEM

### 2.1. Data Preprocessing

The one-second audio segments in our dataset are formatted with a single channel, 44.1kHz sampling rate, and 24-bit resolution per sample. During the audio preprocessing stage, we present the spectrum of the audios in the log-mel domains and apply second-order differencing to the feature maps.

In our work, we transformed audio data into a power spectrogram by using a 2048-length Hanning window and skipping every 1024 samples. From a 1-second audio file, we obtained a spectrum of 44 frames, which was compressed into 256 bins on a mel frequency scale. We also calculated deltas and delta-deltas from the logMel spectrogram and stacked them into the channel axis. To match the number of frames with the delta-delta channel length, we cropped the input feature frames to 36 frames. This resulted in a final shape of [256 × 36 × 3]. To augment the data, we applied random clipping to the temporal dimension of the training set features. In our experiments, the input data with a size of [256 × 36 × 3] was cropped into an input feature map of [256 × 30 × 3] for the training set, and the test set had a dimension of [256 × 36 × 3].

This work was supported by the National Natural Science Foundation of China under Grant No.62276153, and by a grant from the Guoqiang Institute, Tsinghua University. The corresponding author is Wei-Qiang Zhang.

### 2.2 Proposed Model

There were two different basic network architectures we used: MobileNet and Mini-Segnet.

The architecture of MobileNet, which is shown in Figure 1, consists of a total of 7.946K parameters, and all parameters are quantized as int8.

The first component of the network is the frequency split block, which divides the input into low and high frequency components to improve performance. The two components are then processed separately by two convolution blocks. The resulting feature maps are concatenated to form a complete feature map, which is then input into a depthwise convolution block, which is the symbolic module of MobileNet. A dropout layer is used after the GlobalAveragePooling layer to reduce overfitting, and the final predictions are obtained by using the last convolution block.

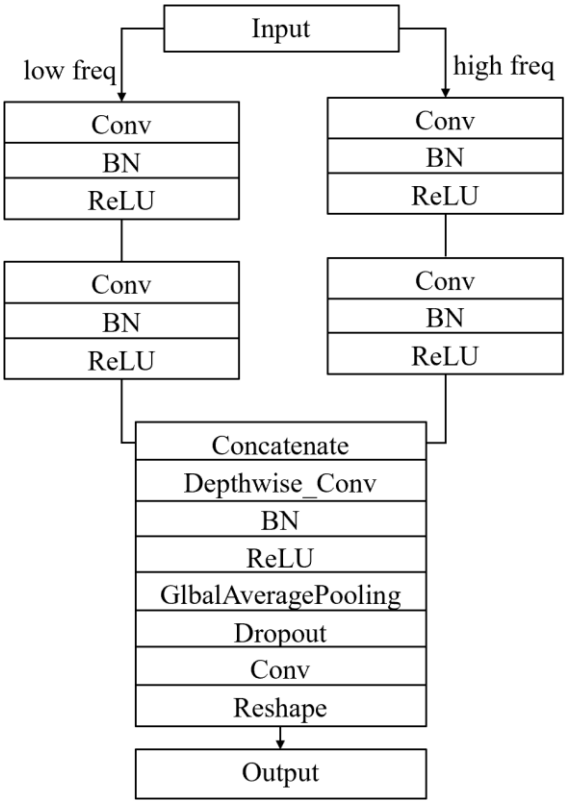


Figure 1: MobileNet architecture

We have also designed a mini-segnet model for low-complexity acoustic scene classification, which is essentially the same as the system we submitted in DCASE2022. It mainly consists of encoder and decoder modules, with ReLU as the non-linear activation function in both.

The encoder module comprises two Conv blocks. The first block contains a convolution layer with a kernel size of 2×3, followed by a mixup normalization layer, a non-linear activation layer, and a max-pooling layer. The second block contains two convolution layers with kernel sizes of 2×3, each followed by a batch normalization layer, a non-linear activation layer, and a max-pooling layer. The output of the encoder is then passed as input to the decoder module.

The decoder module comprises two DeCov blocks as counterparts to the two blocks in the encoder module. In the first DeConv block, up-sampling is performed first, followed by a convolution layer, a batch normalization layer, and ReLU activation. We then repeat these three layers for better performance. The second DeConv block contains four up-sampling layers: 2D convolution, mixup normalization, and ReLU activation.

To avoid overfitting, we include a dropout layer after a dense layer. We utilize two dense layers to output the final predictions. The network architecture is presented in Figure 2.

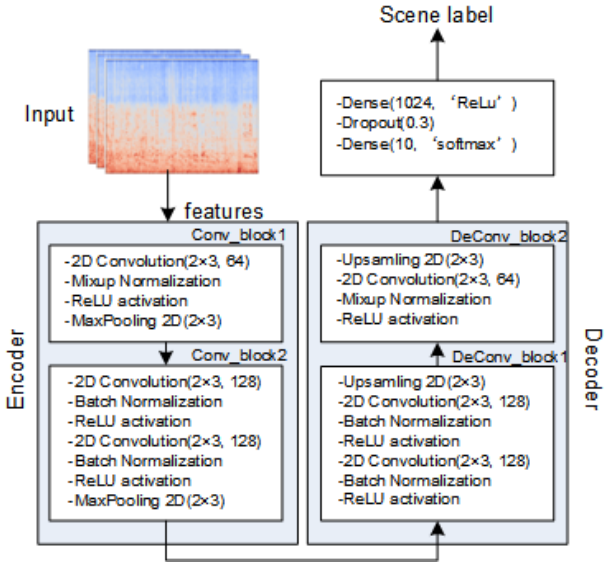


Figure 2: Mini-Segnet architecture

### 2.3 Data Augmentation

To improve the generalization of the model and to prevent overfitting, various data augmentation methods were used.

*2.3.1. Mixup*

Mixup [10] is a simple yet effective data augmentation method that has shown significant improvements in various domains. This technique generates new labeled data samples by linearly combining two pairs of labeled data samples. In our study, we have used mixup with an alpha value of 0.3. Mixup is applied at the mini-batch level, where two data batches along with their corresponding labels are randomly mixed in each training step. The process of mixup involves creating a new training sample (X, y) by mixing two existing training samples (X1, y1) and (X2, y2) using the following equation:

$$X = \lambda X_1 + (1 - \lambda) X_2$$
$$y = \lambda y_1 + (1 - \lambda) y_2$$

where $\lambda \in [0,1]$ is acquired by sampling from the beta distribution $B \in (\alpha, \alpha)$ , and $\alpha$ is a hyper parameters. Besides the data $X_1$ and $X_2$, it is characteristic to mix the labels $y_1$ and $y_2$.

*2.3.2. mask-based method to reduce noise*

We implement a mask-based approach to reduce noise in the time-frequency domain. After extracting the frequency domain features from the audio, we compute a frequency domain mask by comparing the features of the clean audio with the corresponding noise-added frequencies. We then train our model with the noise-added data, using the frequency domain mask as the label. Specifically, we use two frequency masks and a temporal mask with mask parameters of 20 and 3.

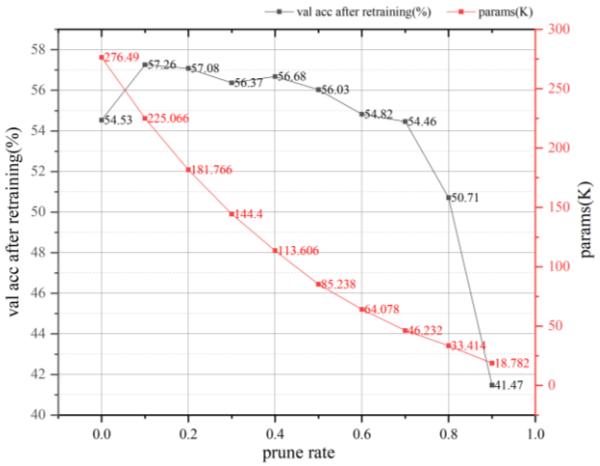


Figure 3: The correlation between validation accuracy, parameters, and pruning rate. Iterative structured pruning is adopted, and retraining is performed after each iteration to fine-tune the model.

Two model compression methods to reduce our model complexity: network pruning and quantization.

*Pruning* is an effective method to produce much smaller, faster, and memory-efficient computational models with minimal loss of accuracy. By training the original SegNet model, which contained 276.49K parameters, and pruning it to 46.232K parameters, the accuracy was reduced by a negligible 0.07%. The correlation between validation accuracy/parameters and pruning rate is shown in Figure 3.

**2.4 Model Compression**

*Model quantization* is a compression technique that converts floating-point values and operations into integer values and operations, respectively. This technique can significantly reduce memory usage without compromising the number of parameters or overall accuracy of the model.

After model compression, our model meets the requirements of MMACs less than 30M and the number of parameters less than 128k, and could achieve a rather small number of parameters, MACs and memory usage with almost no loss of accuracy, which will be elaborated in Section 3.

## 3. EXPERIMENT

**3.1 Experiment setup**

All trainings were performed on GPU using a batch size of 128 and the cross-entropy loss function. In addition, we employed a warm restart [11] learning rate schedule that started with a maximum value of 0.1 after 11.0, 31.0, 71.0, 151.0, and 311.0 epochs, and then decayed according to a cosine pattern to 1×10−5. Each network was trained for 310 epochs. We used different data augmentation methods, such as Mixup with $\alpha = 0.3$ and SpecAugment with a temporal mask and two frequency masks with mask parameters of 3 and 20, respectively, during the training stage for the Mini-SegNet dataset. Our experiments demonstrated that this method can significantly improve the accuracy of acoustic scene classification.

**3.2 Submissions and Results**

The final results on the development set split are reported in Table 1. Below we describe our submissions in detail.

Table 1: the params,model size, MMACs and accuracy of the four submitted systems. The model size could be calculated as [parameter count]*[bit per parameter]/8, with all the parameters quantized to int8, It is numerically equal to the number of the parameter.

| ID | Params (K) | Model size (KB) | MMACs | Train/Val Acc (%) |
|---|---|---|---|---|
| 1 | 7.946 | 7.946 | 3.972 | 63.69/— |
| 2 | 46.232 | 46.232 | 19.466 | 80.19/54.46 |
| 3 | 54.178 | 54.178 | 23.438 | —/56.95 |
| 4 | 15.892 | 15.892 | 7.944 | —/53.18 |

**Submission 1 (MobileNet):** This submission uses a single MobileNet model whose structure is defined in Figure 1. It contains extremely small number of parameters and MMACs. We trained the model with both the train and validation data in development dataset in expected to yield a better result on the evaluation dataset.

**Submission 2 (Segnet):** This submission uses a single SegNet model as demonstrated in Figure 2. The original SegNet model contains more than 276K parameters in total, which exceeds the limitation of Task1, so we applied structured pruning to it. And it decreases to 46.232K parameters when 70% filters were removed.

**Submission 3 (ensemble 1):** This submission is an ensembled model that consists of two models: MobileNet and SegNet. Model ensemble is successful in boosting the system's performance according to previous experiments. We ensemble our models using linear combination as follows:

$$y_{ensemble} = \sum_{n=1}^{N} w_n y_n + b$$

where $N$ is the number of subsystems (N = 2 here), $y_n$ is the output score of each subsystem, $w_n$ is the weight coefficient for each subsys-tem, and $b$ is the bias.

**Submission 4 (ensemble 2):** This submission is also an ensembled model of two MobileNet models which are trained with different inputs. MobileNet 1 is trained with an input of shape [256×36×3], while MobileNet 2 is trained with the features in the time dimension as double the original to improve the accuracy. Both models are evaluated on the test set of dimension [256×36×3].

## 4. CONCLUSIONS

In this report, we describe the methods and techniques used in Task 1 of the DCASE2023 challenge. We extract log-mel spectrograms, delta, and delta-delta features and employ various augmentation techniques. To develop an efficient acoustic scene classification model, we use structured pruned Mini-SegNet and frequency-divided MobileNet. Furthermore, we compress the model by quantizing its parameters. Our system achieves 51.3% accuracy with only 7.946K parameters and 3.972M MACs, and 54.46% accuracy with 46,232K parameters and 19.466M MACs. As the individual models are small, they can also be ensembled to achieve better results within the complexity constraints.

## 5. REFERENCES


[1] Irene Martín-Morató, Francesco Paissan, Alberto Ancilotto, Toni Heittola, Annamaria Mesaros, Elisabetta Farella, Alessio Brutti, and Tuomas Virtanen. Low-complexity acoustic scene classification in dcase 2023 challenge. 2023.

[2] Kim, Byeonggeun and Yang, Seunghan and Kim, Jangho and Chang, Simyung, "QTI Submission to DCASE 2021: Residual Normalization for Device-Imbalanced Acoustic Scene Classification with Efficient Design," DCASE2021 Challenge, Tech. Rep., June 2021

[3] Lee, Joo-Hyun and Choi, Jeong-Hwan and Byun, Pil Moo and Chang, Joon-Hyuk, "Hyu Submission for the DCASE 2022: Efficient Fine-Tuning Method Using Device-Aware Data-Random-Drop for Device-Imbalanced Acoustic Scene Classification," DCASE2022 Challenge, Tech. Rep., June 2022

[4] Morocutti, Tobias and Shalaby, Diaaeldin, "Receptive Field Regularized CNNs with Traditional Audio Augmentations," DCASE2022 Challenge, Tech. Rep., June 2022

[5] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[6] Howard, Andrew, et al. "Searching for mobilenetv3." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[7] Badrinarayanan, V., Kendall, A., Cipolla, R., "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence , 2015.

[8] Xinxin Ma, Yunfei Shao, Yong Ma, Wei-Qiang Zhang. "Deep semantic encoder-decoder network for acoustic scene classification with multiple devices," In: Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, pp. 365–370 , 2020.

[9] Shao, Yun-Fei and Zhang, Xuan and Bing, Ge-Ge and Zhao, Ke-Meng and Xu, Jun-Jie and Ma, Yong and Zhang, Wei-Qiang, "Mini-Segnet for Low-Complexity Acoustic Scene Classification," DCASE2022 Challenge, Tech. Rep., June 2022

[10] H. Zahng, M. Cisse, Y. N. Dauphin, and D. Loped -paz, "mixup: beyond empirical risk minimization," arxiv preprint arxiv:1710.09412, 2017

[11] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with restarts," CoRR, vol. abs/1608.03983, 2016. [online]. Available: http://arxiv.org/abs/1609.03983