ACOUSTIC SCENE CLASSIFICATION BASED ON PRUNED_ GHOSTNET AND FHR _MOBILENET

Technical Report

Lin Zhang, Hongxia Dong

North China University of Technology, Beijing, China
Zl18612608367@163.com

Menglong Wu, Xichang Cai*

North China University of Technology, Beijing, China
wumenglong@126.com

ABSTRACT

This technical report describes our submission for Task 1 of the DCASE2023 challenge. We computed the logarithmic mel spectrogram for each audio segment under the condition of the original sampling rate of 44.1KHz. In addition, to obtain richer feature information, we also computed the first-order and second-order differences on top of the logarithmic mel spectrogram. The resulting spectrogram has 128-frequency bins, 43-time bins, and 3 channels. The feature maps were then fed into classification networks, where we employed two schemes, namely Pruned_GhostNet and FHR_MobileNet.The achieved ac curacies were 47% and 52.8%, respectively, with model para meters of 123.648K and 76.224K, and MACs of 7.375M and 28. 461M.

Index Terms—Acoustic scene classification, GhostNet, FHR_MobileNet, data augmentation.

1. INTRODUCTION

Acoustic Scene Classification (ASC) [1] aims to classify given audio into corresponding scenes through classification models, with ten scene categories included in this task. This year, the ASC task has new changes in the ranking metrics, which not only include accuracy but also involve model parameters and MACs, without including the loss. To adapt to this change, we have selected a new classification model that achieves high accuracy under low complexity conditions.

Unlike last year's challenge, the weight data type of the model can be flexible and does not need to be fixed as INT8; float32 can also be chosen. The maximum number of parameters (including zero values) and MACs are limited to 128K and 30 million, respectively. We selected GhostNet as the neural network model, which was modified based on the model provided in [2] to meet the constraint conditions while ensuring classification performance. Additionally, we explored data augmentation methods [3-5] and attention mechanisms [6-7]. We also adopted the FHR_MobileNet [8], which was used by last year's participants, as the second option, and it also achieved excellent performance.

2. METHOD

In this section, we describe the design of the feature extraction and classification model based on the baseline model.

A. Acoustic Features

We aim to select mature and perceptually meaningful feature representations. According to human auditory features, we are more sensitive to differences in low frequencies, and the human ear cannot perceive frequency linearly. Therefore, we first consider using the log Mel spectrogram feature. As a commonly used audio feature extraction method, the log Mel spectrogram contains time and frequency domain information and perceptually relevant amplitude information, and its core lies in the Mel scale, which is more consistent with the auditory characteristics of the human ear.

In addition, since speech signals are temporally continuous, the feature information extracted by frame-by-frame analysis only reflects the characteristics of the current frame of the speech signal. To better reflect the temporal continuity of the feature, the dimensions of the front and back frame information can be added to the feature dimension, commonly used are the first-order difference and the second-order difference. Therefore, we perform first-order and second-order differences on the log Mel spectrogram, and cascade them along the channel dimension as the input to the neural network, thereby obtaining richer feature information and ensuring a more comprehensive understanding of the data.

B. Model Design

The GhostNet used in this paper is modified based on the original paper. The original GhostNet contains 16 G-benck modules, to reduce parameter and computational costs, only the first 6 G-benck modules are selected, and the 1×1 convolution layer before the fully connected layer is removed. At the same time, the output channel number of the penultimate 1×1 convolution layer is reduced from 960 to 160. The specific structure is shown in Table 1. The FHR_MobileNet structure used in the second solution is described in [8]..

Table1: Architecture of Pruned_GhostNet.

Operator	#exp	#out	Stride
Conv2d 3×3	-	16	2
G-bneck	16	16	1
G-bneck	48	24	2
G-bneck	72	24	1
G-bneck	72	40	2
G-bneck	120	40	1
G-bneck	240	80	2
Conv2d 1×1	-	160	1
GlobalAveragePooling2d	-	-	-
Dense	-	10	=

3. EXPERIMENT

3.1 Training Setup

All experiments in this paper were conducted on the development dataset of DCASE2023 Task1 [9] (TAU Urban Acoustic Scenes 2022 Mobile, development dataset). The dataset consists of 230,359 audio clips, each with a duration of 1 second, collected from 9 devices: 3 real devices (A, B, C) and 6 simulated devices (S1~S6). The dataset includes 10 urban acoustic scenes, namely the airport, shopping mall, metro station, pedestrian street, public square, urban traffic, tram, bus, metro, and park.

For feature extraction, 128 Mel filters were used to process the audio signals, and fast Fourier transform (FFT) was performed on them with a window length and frame length of 0.04 seconds and 0.02 seconds, respectively, resulting in a spectrogram of size 128×43. Then, first-order and second-order differential features were extracted, and the three feature spectrograms were concatenated along the channel dimension to obtain an input feature spectrogram of size 128×43×3. During the training phase, the batch size was set to 128, and the training iteration was set to 200. In addition, Mixup and SpecAugment data augmentation techniques were employed to optimize the training process.

3.2 Experiment Results

The experiments were conducted on an NVIDIA RTX2080 Ti GPU using the TensorFlow and Keras frameworks on the Windows 10 operating system. Both of our submitted systems outperformed the baseline in terms of accuracy, as shown in Table 2. Neither of our systems employed TFLite quantization. The GhostNet model achieved a classification accuracy of 47%, with 123.648K parameters and 7.375M MACs. The submitted FHR_MobileNet achieved an accuracy of 52.8%, with 76.224K parameters and 28.461M MACs.

4. CONCLUSION

In this technical report, we not only used the model that we submitted last year but also modified the GhostNet model to better meet the competition criteria. Additionally, we employed two data augmentation techniques, Mixup and Specaugment, to prevent overfitting. Ultimately, both models achieved accuracies of 47% and 52.8%, respectively, within the required complexity constraints.

Table2: Results of different models.

Scene	baseline	GhostNet	FHR_MobileNet
Airport	39.4%	47.2%	40.8%
Bus	29.3%	44.9%	57.0%
Metro	47.9%	49.9%	55.2%
Metro_station	36.0%	32.0%	43.6%
Park	58.9%	74.8%	81.6%
Public_square	20.8%	22.1%	31.6%
Shopping_mall	51.4%	47.5%	62.6%
Street_ pedestrian	30.1%	31.8%	31.7%
Street_traffic	70.6%	64.6%	72.9%
Tram	44.6%	54.7%	50.4%
Average	**42.9%**	**47.0%**	**52.8%**
Model_size	**46.512K**	**123.648K**	**76.224K**
MACs	**29.23M**	**7.375M**	**28.461M**

5. REFERENCES

[1] Martín-Morató I, Paissan F, Ancilotto A, et al. Low-complexity acoustic scene classification in DCASE 2022 Challenge[J]. arXiv preprint arXiv:2206.03835, 2022.

[2] Han K, Wang Y, Tian Q, et al. Ghostnet: More features from cheap operations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1580-1589.

[3] Li Y, Cao W, Xie W, et al. Low-Complexity Acoustic Scene Classification Using Data Augmentation and Lightweight ResNet[C]//2022 16th IEEE International Conference on Signal Processing (ICSP). IEEE, 2022, 1: 41-45.

[4] Zhang H, Cisse M, Dauphin Y N, et al. Mixup: Beyond empirical risk minimization[J]. arXiv preprint arXiv:1710.09412, 2017.

[5] Park D S, Chan W, Zhang Y, et al. Specaugment: A simple data augmentation method for automatic speech recognition[J]. arXiv preprint arXiv:1904.08779, 2019.

[6] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF

conference on computer vision and pattern recognition. 2021: 13713-13722.
[7] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
[8] Hongxia Dong, Lin Zhang, Xichang Cai, et al. Acoustic Scene Classification Based on Fhr_mobilenet.2022.
[9] Mesaros A, Heittola T, Virtanen T. A multi-device dataset for urban acoustic scene classification[J]. arXiv preprint arXiv:1807.09840, 2018.