ACOUSTIC SCENE CLASSIFICATION BASED ON FEATURE FUSION AND DILATED-CONVOLUTION

Technical Report

Junfei Yu, Runyu Shi, Tianrui He, Kaibin Guo*

Multimedia Technology Department
Xiaomi INC.
Beijing 100085, CHN
yujunfei@xiaomi.com

ABSTRACT

This technical report describes our submission for Task 1 of the DCASE Challenge 2022. The goal of task 1 is to classify the recorded audios for acoustic scene classification using an int8 quantized model that does not exceed 128KB in size. In our submission, a variety of time-frequency features are extracted and fused to be the input of the deep learning network. As the backbone of the network, the dilated-convolution is applied for embedding of various input features. Furthermore, we make use of multiple time-frequency data augmentation on the original data to increase the diversity of the data. After the network training is completed, the variable type of the weight data is converted into INT8. This INT8 model achieves a log loss of 1.305 and an accuracy of 51.7% on the standard test set of the TAU Urban Acoustic Scenes 2022 Mobile development dataset.

Index Terms— Acoustic scene classification, dilated-convolution, time-frequency data augmentation

1. INTRODUCTION

Acoustic Scene Classification (ASC) aims to identify sound scenes of the recorded audios, the recent research on which has been led by the DCASE community. In the past few years, a large number of ASC methods have been released, such as Resnet-based models[1], two-stage-based models[2], etc. In DCASE2022, Task 1 is designed as follows: given a one-second-long audio, submit a method to recognize the scene of this audio (e.g., airport, park, street, etc.), subject to the constraints of MACC and model size.

In order to make full use of the time-frequency characteristics of audio data, we adopted various methods for data processing in data enhancement and feature extraction, and designed a suitable CNN network model for embedding. First of all, various data augmentation methods in the time-frequency domain are applied to increase the diversity of training data and enhance the generalizability of the model. Next, the log amplitudes of the Mel (log-mel) spectrogram, Spectral Entropy(SE) and Spectral Flatness(SF) are extracted from origin audio data as the input of our model. This report is structured in four sections. Section 2 our experiment setup will be discussed. Results will be shown in Section 3. Section 4 is references.

2. EXPERIMENT SETUP

2.1. Feature Extraction

We adopted the TAU Urban Acoustic Scenes 2022 Mobile Development dataset as training and validation sets. 128 log-mel spectrogram is calculated under the sampling rate of 44.1KHz for each audio slice (20ms) with 50% overlap. As effective features of audio signals, SE and SF are constantly used for audio classification tasks such as speech recognition[3]. In our report, SE and SF are computed with each audio data as input by the LibROSA library[4]. Thus, the size of log-mel spectrogram is 128*101, the size of SE and SF are both 1*101.

2.2. Data Augmentation

To improve the performance and generalization ability of the model, we implemented data augmentation in the time and frequency domains respectively. The applied time-domain data augmentation includes: a) mixup of multiple signals; b) signal data plus random noise; c) random time shift of the signal data. The frequency-domain data augmentation includes: SpecAugment[5] and SpecCorrection[6]. After data augmentation, the amount of data has doubled to about 200,000. Besides, in order to enable the model can fit to all kind of devices, a device-wise data balance strategy has been utilized. In final, the amount of training data is 160,000.

2.3. Backbone

We trained a 5-layer multi-input CNN-based model as the backbone. For the three input features, the convolutional neural network is used for embedding. Dilated-convolution is used in the first two layers of the network to increase the receptive field. Each convolutional layer is followed by a Batch Normaliza-

* Thanks to Xiaomi Inc. for funding.

tion(BN) operation and ReLU6 activation layer. Max-pooling and Global Average Pooling(GAP) are used to reduce the data dimension. At the end of our network, the three-way tensors are concatenated. Finally, softmax is used to get classification results. Figure 1 shows the backbone architecture used in this submission.

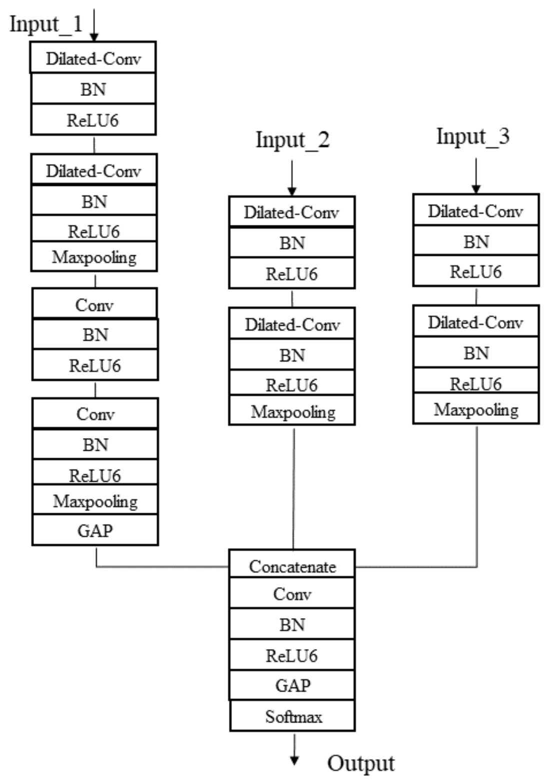

Figure 1: The backbone architecture.

2.4. Training

For training the model, we make use of backpropagation and stochastic gradient descent with a batch size of 128 and the cross-entropy loss function. In training, a learning rate decay strategy is used, which will reduce the learning rate to 80% when the accuracy of the validation set does not decrease for 10 consecutive epochs.

2.5. Quantization and Inference

By applying post-training quantization, we convert the data type of the weights in the model to INT8, which reduces the model size but decrease the accuracy. Results were obtained by using the quantized model.

3. RESULTS

Table 1 shows the accuracy of each categories and the overall Accuracy on the test dataset of the DCASE 2022 development dataset.

Table 1: Results of Task 1

Categories	**Baseline**		**Our method**	
	Accuracy	logloss	Accuracy	logloss
Airport	39.4%	1.534	51.0%	1.326
Bus	29.3%	1.758	58.6%	1.121
Metro	47.9%	1.382	40.0%	1.510
Metro_station	36.0%	1.672	39.1%	1.711
Park	58.9%	1.448	76.7%	0.847
Public_square	20.8%	2.265	38.3%	1.717
Shopping_mall	51.4%	1.385	57.6%	1.137
Street_pedestrian	30.1%	1.822	27.7%	1.699
Street_traffic	70.6%	1.025	74.4%	0.899
Tram	44.6%	1.462	47.9%	1.231
Overall	42.9%	1.575	51.7%	1.305

4. REFERENCES

[1] Hu H, Yang C H H, Xia X, et al. A two-stage approach to device-robust acoustic scene classification[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 845-849.

[2] Byttebier L, Desplanques B, Thienpondt J, et al. Small-footprint acoustic scene classification through 8-bit quantization-aware training and pruning of ResNet models[R]. DCASE2021 Challenge, Tech. Rep, 2021.

[3] Toh A M, Togneri R, Nordholm S. Spectral entropy as speech features for speech recognition[J]. Proceedings of PEECS, 2005, 1: 92.

[4] https://librosa.github.io/librosa/.

[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] Nguyen T, Pernkopf F, Kosmider M. Acoustic scene classification for mismatched recording devices using heated-up softmax and spectrum correction[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 126-130.