SELF-ENSEMBLE WITH MULTI-TASK LEARNING FOR LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION

Technical Report

Reiko Sugahara , Ryo Sato , Masatoshi Osawa , Yuuki Yuno , Chiho Haruta

RION CO., LTD.
3-20-41 Higashimotomachi, Kokubunji, Tokyo, Japan
{r-sugahara, sato.ryou, m-osawa, y-yuno, c-haruta}@rion.co.jp

ABSTRACT

This technical report describes a procedure for Task 1 in Detection and Classification of Acoustic Scenes and Events (DCASE) 2022. The proposed method adopts MobileNet-based models with log-mel energies and deltas as inputs. The accuracy was improved by self-ensemble with multi-task learning. Data augmentations, e.g., mixup, SpecAugment, and spectrum modulation, were applied to prevent overfitting. To meet system complexity requirements, we adopted depth-separable convolution and quantization aware training. The model contains 120,505 parameters and requires 26.607 million multiply-and-accumulate operations. Consequently, the proposed system achieved a 56.5% accuracy and a log-loss of 1.179 based on the development data.

Index Terms— acoustic scene classification, multi-task learning, quantization aware training

1. INTRODUCTION

Task 1 in DCASE 2022 denotes the acoustic scene classification (ASC) task [1][2], which has been the main task of DCASE since its inception. The task rules are getting harder every year to accommodate the real world. At present, we need to address low-complexity and an audio dataset, which is multiple devices and short-time data length.

- Low-complexity: The task applies model complexity limits. First, the maximum number of parameters is 128K with INT8 and includes zero-valued ones. Second, the maximum number of million multiply-accumulate operations (MMAC) per inference is 30. The limits are modeled after Cortex-M4 devices that are used for Arduino Nano 33@64MHz.
- Audio dataset: The dataset of this task is TAU Urban Acoustic Scenes 2022 Mobile, development dataset[3], with a 44.1 kHz sampling rate, 24-bit resolution, and ten types of scenes. The development dataset comprises the information of ten cities recorded using nine devices. However, the evaluation dataset comprises twelve cities recorded using eleven devices, and two cities and five devices are not available in the development set. Moreover, the length of each data point is only 1 s. Therefore, we need to architect models that can be inferred from limited information and are robust to multiple devices.

First, we describe the method for the preprocessed signal (Section 2.1). Subsequently, we explain the network architecture, which is based on MobileNet (Section 2.2) and multi-task learning (Section 2.3). After introducing data augmentation (Section 3) and model compression (Section 4), we state the experimental results (Section 5) and conclude the report (Section 6).

2. PROPOSED SYSTEMS

2.1. Audio Signal Preprocessing

We adopt log-mel energies and deltas as inputs at 44.1 kHz. We have tried several sampling frequencies, and the result is better without downsampling. Log-mel energies exhibit a 256-dimensional structure with a window length and frameshift of 4096 and 1024, respectively. As the length of each data point is 1 s and deltas are added, the input size of our models is [2,256,44]. The data are standardized after data augmentation, which is described in Section 3.

2.2. Network Architecture

We build a model based on MobileNet[5], which is known for low-complexity architecture by depthwise separable convolution (DSC) [6][7]. The outputs of the first convolution layer are split along frequency dimension and concatenated after the DSC. Moreover, we adopt two residual connections[8]. Figure 1 shows the structure of the model.

2.3. Self-ensemble with Multi-Task Learning

We added extra labels to each data point for multi-task learning (MTL). The additional labels are “indoor,” “outdoor,” and “transportation,” citing DCASE 2020 Task 1. Table 1 lists the correspondence between the ten class of this task and three additional class assigned. Then, we apply MTL with projecting conflicting gradients (PCGrad)[9]. PCGrad projects a gradient for smooth training when there is a gradient conflict between two tasks. In addition, the outputs of the three-class classification have been extended to ten class (Figure 2). The ensemble is executed using the outputs of the original ten-class and extended ten classes (Figure 3). If ensembled with high-perfomance three-class classification, the output of unlabeled class will be smaller. Then, the log loss will be lower.

3. DATA AUGMENTATION

To prevent overfitting and improve robustness, we have adopted several data augmentation (DA) methods. These methods are performed in the time-frequency domain during training.

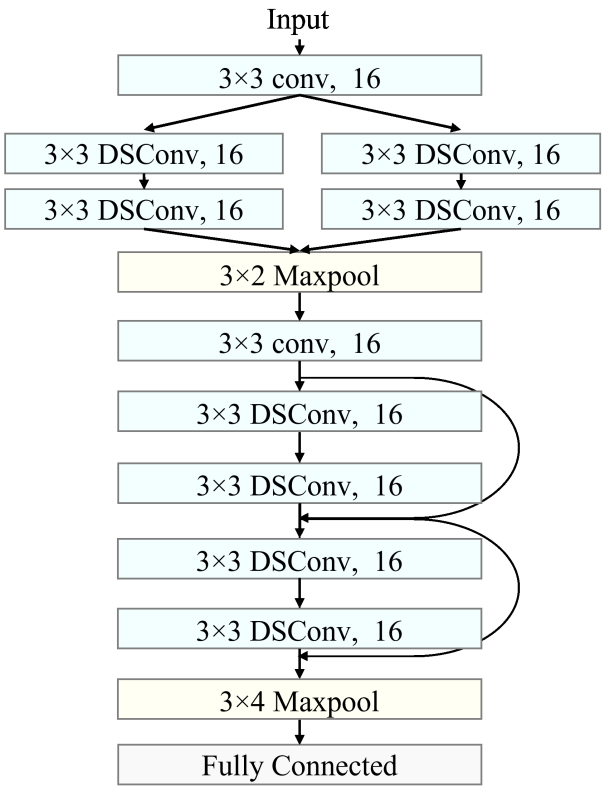


Figure 1: Structure of the model.

- Mixup[11]: Mixup is the process of mixing two sound sources in an arbitrary proportion. Herein, we set $\alpha = 0.2$, which is a parameter of β-distribution. In addition, referring to the method of Hao et al.,[12], Mixup is performed with overall data up to 60 epochs and half of the data after that.
- SpecAugment[13]: SpecAugment is a commonly used DA technique in ASC, which includes functional warping, frequency channel masking blocks, and timestep masking blocks. We apply two masking lines for each dimension, and the maximum thickness of one line is 2.
- Spectrum modulation: As the spectrum modulation was confirmed to be very effective in our submission of the DCASE 2021 challenge[4], the same method was used this year. Most of the provided datasets were recorded using device A; therefore, the data are imbalanced. We dealt with this problem by applying a frequency energy difference to the data of non-device A. Figure 4 shows spectrum modulated data.

4. MODEL COMPRESSION

According to DCASE 2022, the submitted model should be INT 8; however, we can train models with float32 in PyTorch. Therefore, we adopted quantization aware training (QAT) as the quantization method for our models. QAT optimizes a model that can consider quantization errors during training. It can quantize a model without degrading performance than post-training quantization. In addition, we performed knowledge distillation (KD) to learn the outputs of

Table 1: Relationship between three and ten-class classification.

	10 classes of this task	3 additional classes
1	Airport	Indoor
2	Bus	Transport
3	Metro	Transport
4	Metro station	Indoor
5	Park	Outdoor
6	Public square	Outdoor
7	Shopping mall	Indoor
8	Street pedestrian	Outdoor
9	Street traffic	Outdoor
10	Tram	Transport

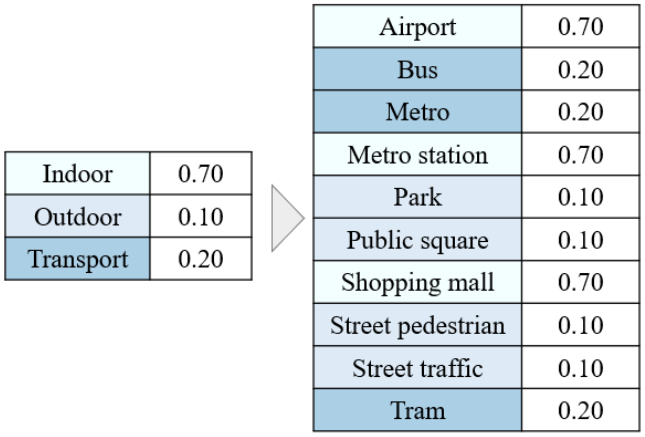


Figure 2: Extension of three additional class to ten class.

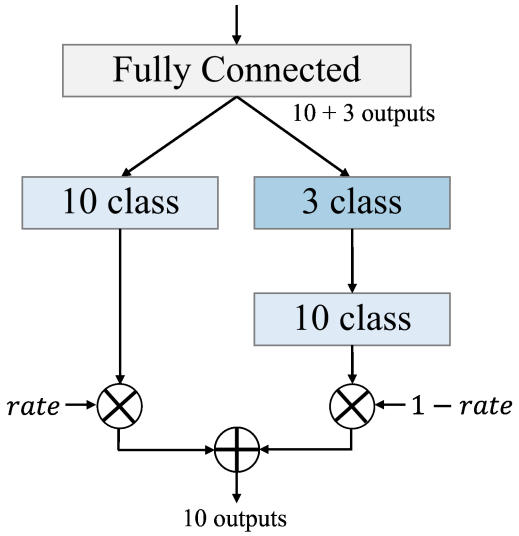
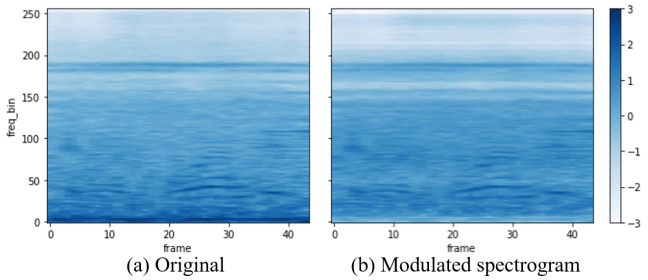


Figure 3: System of self-ensemble with MTL.



(a) Original (b) Modulated spectrogram

Figure 4: Log-mel energies as input; (a) Original and (b) Modulated spectrum.

the large model and improve the accuracy of the small model[10]. This method was not compatible with MTL;therefore, we could not perform it simultaneously with KD and MTL.

5. RESULTS

We report the performance of the submitted model based on the development set. In the challenge rule, we can submit four systems and will be ranked by macro-average multiclass cross-entropy (log loss) (average of the class-wise log loss). We used all the development datasets to train the submitted model;however, this report shows the results of validation data when training with the split used in the baseline of DCASE 2022. It implies that the training data do not include validation and are not used for spectral modulation;therefore, we can confirm the robustness of our models.

Table 2 lists the conditions and results of the submissions. Submissions 1, 2, and 3 are self-ensembled with MTL, and Submission 4 is a KD model. Weighted score averages are used while self-ensembling and the rates are also listed in Table 2. The DCASE 2022 task1 baseline system exhibits a log loss of 1.575 and an accuracy of 42.9 % in terms of performance. However, a log loss of 1.179 for Submission 3 and an accuracy of 57.1% for Submission 4 have been reported with regard to submission performance. Figure 5 shows the accuracy of Submission 2 for each class. Although Bus, Park, and Street Traffic exhibit high accuracy, Street pedestrian is not very accurate. This has not been improved, and can be considered in future studies.

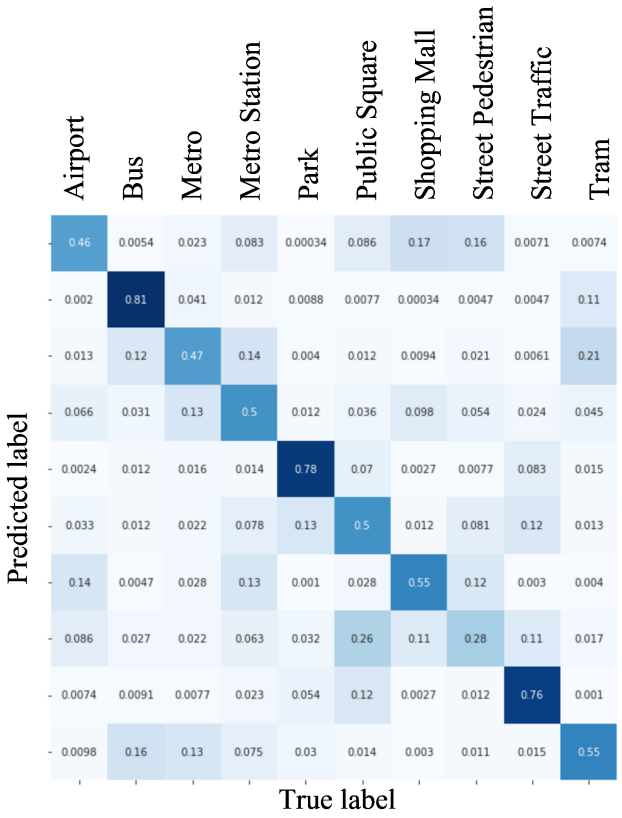


Figure 5: Accuracy confusion matrix for the validation data of Submission 2.

6. CONCLUSION

In this technical report, we described a system for the low-complexity acoustic scene classification Task 1 of DCASE challenge 2022. The network architecture is based on the self-ensemble of MobileNet-based models with MTL. We tried to improve the performance of the submission models by applying DA, QAT, and KD, even though the number of calculations was limited. Thus, the accuracy of the submitted model is 14% higher than the baseline with regard to the development dataset.

7. REFERENCES

[1] https://dcase.community/challenge2022/index

[2] 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 2022 challenge” arXiv:2206.03835, 2022.

[3] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen, “Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions” arXiv:2005.14623, 2020.

[4] Reiko Sugahara, Masatoshi Osawa, and Ryo Sato, “Ensemble of simple resnets with various mel-spectrum time-frequency resolutions for acoustic scene classification” 2021.

[5] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto and Hartwig Adam, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications” arXiv:1704.04861, 2017.

[6] Laurent Sifre, “Rigid-motion scattering for image classification” 2014 Ph.D.thesis.

[7] Francois Chollet, “Xception: Deep Learning With Depthwise Separable Convolutions” arXiv:1610.02357, 2017.

[8] K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition” arXiv:1512.03385, 2015.

[9] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman and Chelsea Finn, “Gradient Surgery for Multi-Task Learning” arXiv:2001.06782, 2020.

Table 2: Methods and results with regard to the development dataset for each system.

	MTL	rate	KD	DA	QAT	Epoch	Params	MMSC	Acc	Loss
1	✓	0.8	-	✓	✓	120	120K	26.6	56.25%	1.1988
2	✓	0.8	-	✓	✓	100	120K	26.6	56.49%	1.1790
3	✓	0.7	-	✓	✓	100	120K	26.6	56.55%	1.1817
4	-		✓	✓	✓	100	123K	26.6	57.14%	1.2142

[10] Jangho Kim, Minsung Hyun, Inseop Chung and Nojun Kwak, “Feature Fusion for Online Mutual Knowledge Distillation” arXiv:1904.09058, 2019.

[11] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin and David Lopez-Paz, “mixup: Beyond Empirical Risk Minimization” arXiv:1710.09412, 2017.

[12] Hao Yu, Huanyu Wang and Jianxin Wu, “Mixup Without Hesitation” arXiv:2101.04342. 2021.

[13] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk and Quoc V. Le, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition” arXiv:1904.08779, 2019.