ENSEMBLE OF SIMPLE RESNETS WITH VARIOUS MEL-SPECTRUM TIME-FREQUENCY RESOLUTIONS FOR ACOUSTIC SCENE CLASSIFICATIONS

Technical Report

Reiko Sugahara , Masatoshi Osawa , Ryo Sato

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

ABSTRACT

This technical report describes procedure for Task 1A in DCASE 2021[1][2]. Our method adopts ResNet-based models with a mel spectrogram as input. The accuracy was improved by the ensemble of ResNet-based simple models with various mel-spectrum time-frequency resolution. Data augmentations such as mixup, SpecAugment, time-shifting, and spectrum modulate were applied to prevent overfitting. The size of the model was reduced by quantization and pruning. Accordingly, the accuracy of our system was achived 70.1% with 95 KB for the development set.

Index Terms— acoustic scene classification, ResNet, data augmentation, ensemble, pruning

1. INTRODUCTION

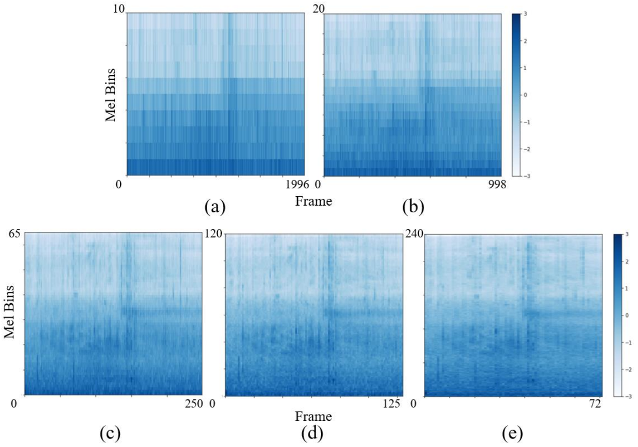
Task1 SubtaskA (Task 1A) attempts to realize an acoustic scene classification that is robust to multiple devices and has a limited model size. The development dataset comprises ten cities recorded by nine devices. In addition, we analyze ten types of scenes. Because five new devices were added to the test data, it is necessary to create a system that is robust to the differences between the devices. The task also targets low complexity solutions for the classification problem in terms of model size. The data adopted must be 128 KB or less. The dataset for this task is TAU Audio-Visual Urban Scenes 2020[3], with a 44.1 kHz sampling rate and 24-bit resolution.

In this report, we first explain the method for the preprocessed signal. We adopt log-mel power as input. Input sizes are prepared with various resolutions in the time and frequency for ensemble learning (Section 2.1). Subsequently, we describe network architecture, which is based on ResNet and has a model size of 67.7 KB before compression. Three models with different inputs are prepared and ensemble learning is executed (Section 2.2). After introducing data augmentations (Section 3) and model size reduction (Section 4) , we provide some experimental results (Section 5) and conclude the report (Section 6).

2. PROPOSED SYSTEMS

2.1. Audio Signal Preprocessing

We adopt log-mel power as input for the ResNet-based model. The data set is mono and the common sampling rate is 44.1 kHz. We attempted to improve the accuracy via ensemble learning using several models with different feature values. Input sizes are prepared with various resolutions in the time and frequency, which is for an ensemble of models that specialize in different acoustic features. We prepared five types of inputs as shown in Figure 1.



(a)

(b)

(c)

(d)

(e)

Figure 1: Log-mel power as input; each sizes are 10 × 1996 (a), 20 × 998 (b), 65 × 250 (c), 120 × 125 (d), and 240 × 72 (e).

2.2. Network Architecture

ResNet[4] is a type of neural network known for its high performance, which has deep layers and inputs to some layers passed directly or as shortcuts to other layers. However, the base ResNet is a very large model and unsuitable for Task 1A; hence we use a simple ResNet with fewer layers as shown in Figure 2.

An ensemble is a fusion of different independently trained models, which is known for its significant contribution to the improvement of accuracy. Therefore our network is architected by an ensemble of models with different inputs. Using classified outputs of each ResNet-based model as input, we designed a fully connected model, as shown in Figure 3.

3. DATA AUGMENTATION

To prevent overfitting and accommodate device sources in the test data, we adopt several data augmentation strategies. All strategies do not generate extra training data.

- mixup[5]: Mixup is the process of mixing two sound sources

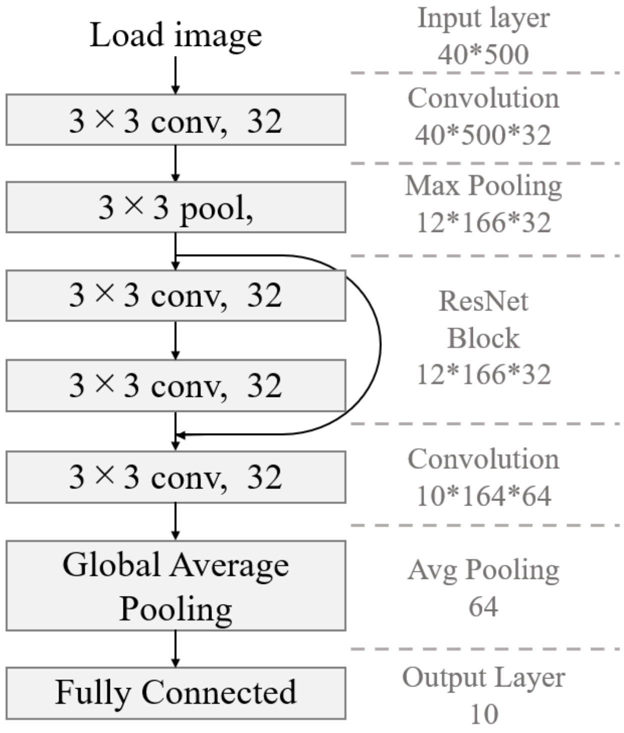


Figure 2: simple ResNet structure.

in an arbitrary proportion. Here we set α = 1.3, which is a parameter of β-distribution.

- SpecAugment[6]: SpecAugment involves warping the features, masking blocks of frequency channels, and masking blocks of time steps. However, in our system, masking blocks of frequency channels is solely applied. The masking rate is limited from 0% to 50%.
- time-shifting: An index is determined randomly from the time axis, and the data is shifted from there.
- spectrum modulate: The above three methods are well-known and conventional methods. To further improve robustness for unknown devices, we took original strategies based on a previous study[7]. Most of the provided datasets are recorded with device A. Therefore, we generated 3747 kinds of characteristics that can simulate devices other than A, which are generated by data obtained with device (B or C or S1 or S2 or S3) - A in log-mel power. By introducing the characteristics randomly for each batch, we aimed to cover the characteristics of the devices that were not in the original data set.

4. MODEL COMPRESSION

In the DCASE CHALLENGE 2021 Task1A, the model complexity limit is set to 128 KB for non-zero parameters. To compress our model size, we applied pruning and quantization to the model, which is trained with float32.

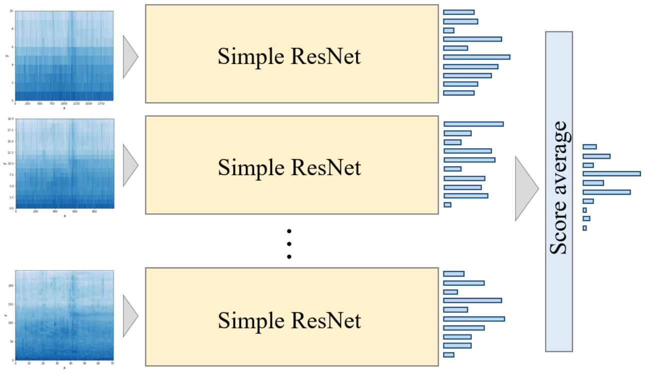


Figure 3: Example of a diagram with experimental results.

4.1. Pruning

Pruning is one of the methods for compressing the model size by changing unimportant weights to zero. The pruning method, which we applied was provided by Tensorflow Lite[8]. To maintain the classification accuracy of the original model, as much as possible, we set the final sparsity to 75%.

4.2. Quantization

As an alternative method for compressing the model size, we applied post-trainingquantization provided by Tensorflow lite to our model. The weights of the original model are float32. By quantizing them to int8, the size of the model can be compressed to approximately 1/4 of its original size.

5. RESULTS

The dataset of TAU Urban Acoustic Scenes 2020 Mobile Development dataset, it contains three real devices (A, B, C) and six simulated devices (S1–S6) . Some devices (S4, S5, S6) are solely included in the test subset. We report the performance of our system compared to the baseline on the development set. The systems will be ranked by macro-average multiclass cross-entropy (Log loss) (average of the class-wise log loss).

Table 1 presents the results of our systems and the baseline in each device. The system 1, 2 and 4 are the ensemble of five models and the system 3 is the ensemble of three models. We adopt the weighted score average for the system 1 and 3, the score average for the system 2 and the weighted score average by the dance layer for the system 4 as the decision making method. Compared with the DCASE 2021 task1A baseline system, our network structures are improved by approximately 20%. The models worked each devices evenly, without being weak in any specific device. From Figure 4, we can observe that the highest accuracy among all the classes is of the park at 96%.

6. CONCLUSION

In this technical report, we described a system for acoustic scene classification Task 1A of DCASE challenge 2021. The network architecture is an ensemble of ResNet-based simple models with various mel-spectrum time-frequency resolutions. During training, data augmentation was applied to prevent overfitting and improve

Table 1: Log loss on the development dataset for each systems.

Model Name	Device-wise log-loss									Log loss	Accuracy [%]
	A	B	C	S1	S2	S3	S4	S5	S6		
Baseline	1.018	1.294	1.282	1.552	1.441	1.421	1.934	1.583	1.729	1.473	47.7
Proposed model 1	0.910	0.973	0.891	0.964	0.971	0.904	1.040	0.991	0.981	0.9585	70.1
Proposed model 2	0.896	0.982	0.898	0.991	1.006	0.931	1.053	1.009	1.007	0.9747	69.7
Proposed model 3	0.934	0.956	0.865	0.919	0.899	0.864	1.021	0.993	0.986	0.9373	66.8
Proposed model 4	1.054	1.125	1.061	0.988	1.082	0.894	1.270	1.111	0.972	1.062	68.8

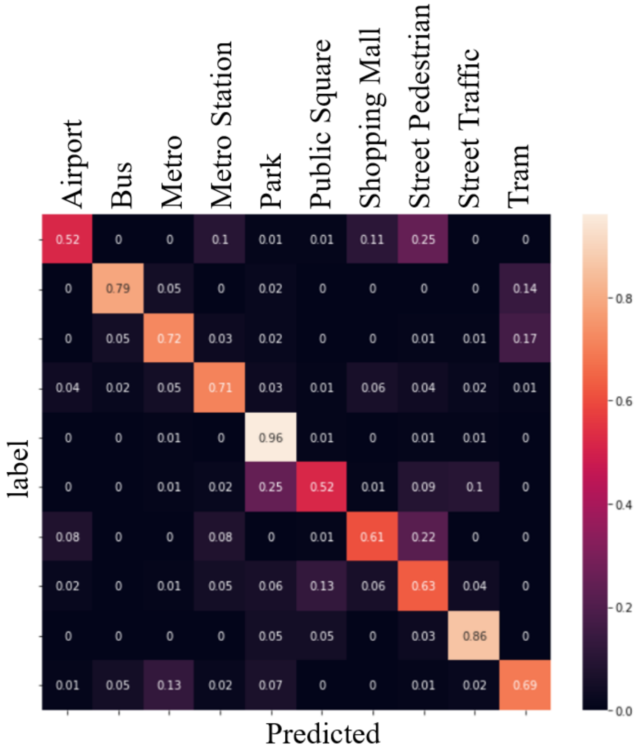


Figure 4: Confusion matrix for the validation data of system 1.

robustness to multiple devices. This allowed our proposal model to perform better than the baseline.

7. REFERENCES

[1] http://dcase.community/challenge2021/index

[2] Irene Martín-Morató, Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. Low-complexity acoustic scene classification for multi-device audio: analysis of dcase 2021 challenge systems. 2021. arXiv:2105.13734.

[3] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020). 2020. Submitted. URL: https://arxiv.org/abs/2005.14623.

[4] K. He, X. Zhang, S. Ren, J. Sun, " Deep Residual Learning for Image Recognition, " arXiv preprint arXiv:1512.03385, 2015.

[5] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," arXiv preprint arXiv:1710.09412, 2017.

[6] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition , " arXiv preprint arXiv:1904.08779, 2019.

[7] M. Kosmider, " Spectrum Correction: Acoustic Scene Classification with Mismatched Recording Devices," INTERSPEECH, pp. 4641–4645, Jan. 2020.

[8] https://www.tensorflow.org/lite/performance/post_training_quant