LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION USING DEEP SPACE SEPARABLE DISTILLATION AND MULTI-TASK LEARNING

Technical Report

Kangli Wang, Yiling Wu, Yanxiong Li

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China
eewkl@mail.scut.edu.cn, 202030242140@mail.scut.edu.cn, eeyxli@scut.edu.cn

ABSTRACT

This technical report describes our system for Task 1 in Detection and Classification of Acoustic Scenes and Events (DCASE) 2023. We propose a deep space separable distillation block as the basic unit of the model, using its strong block processing ability to continuously cut the high-frequency and low-frequency parts of the log-Mel spectrogram. The accuracy is improved by multi-scale embedding and multi-task learning methods. To prevent overfitting, we adopt data augmentation methods such as mixing, speculation and spectral modulation. Quantization aware training is adopted to quantize the model to meet the requirements of edge devices with low complexity constraints. The proposed system achieves a 53.3% accuracy on the development dataset with only a parameter count of 45.16 kB and the MACs of 8.64 M[1].

***Index Terms*—**Acoustic scene classification, Deep space separable distillation block, Multi-task learning, Quantization aware training

1. INTRODUCTION

Task 1 of DCASE2023 is acoustic scene classification, which is a classic DCASE task. In order to simulate the real world, the rules of the task are changing every year. Now, we need to develop a system with a few parameters, low MACs and high accuracy. This task has the following specific constraints:

- generalization across different recording devices and cities.
- For low complexity requirements in terms of the number of parameters (128 kB), with using the Int8 type for forward inference, and a limited number of multiply-accumulate operations (30 million MACs).
- Systems will be ranked by a combination of different criteria:

$$R = 0.5 * R_{ACC} + 0.25 * R_{MEM} + 0.25 * R_{MAC} \quad (1)$$

In recent years, CNN and ResNet have been widely used in acoustic scene classification, and attention has further improved the accuracy of classification [2-6]. And multi-task learning [7] was also applied to acoustic scene classification, and it has better results on the TAU Urban Acoustic Scenes 2022 Mobile dataset [8]. Deep separable convolution [9] also achieved good results on 2020 ASC task 1. Li et al [10] proposed a structure of distillation blocks, which can reduce the size of the model and MACs at the same time, and has achieved very good results in the field of image super-resolution.

In this report, we propose a Deep Separable Multiscale Networks (DSMN) which uses the distillation block to process log-Mel spectrogram. We constructed the basic unit Deep Space Separable Convolution (DSSC), and then built the basic module Deep Space Separable Distillation Block (DSSDB) based on the basic unit. Then the basic modules are stacked to get our final model DSMN. Finally, the model is quantized to 8 bit.

2. DATA PREPROCESSING AND AUGMENTATION

2.1. Audio Preprocessing

All audio recordings are formatted with a mono channel and 44.1 kHz sampling rate. For each recording, an input feature is extracted by using a Short Time Fourier Transformation (STFT) with Hanning window, whose length is 4096 and overlap is 25%.

Next, log-Mel filter banks end up with 256 frequency bins are performed for producing log-Mel spectrogram. In addition, the deltas coefficient of the log-Mel spectrogram are calculated and stacked into the channel axis. Therefore, the final input size of our models is $2 \times 256 \times 44$, where 2, 256 and 44 represent channels, numbers of frequency-band and frame, respectively.

2.2. Data Augmentation

Data augmentation(DA) is an effective way to prevent overfitting and improve the generalization ability of our networks. Inspired by [11-13], we have adopted several data augmentation (DA) methods performed in the time-frequency. The methods used in our system are described as follows.

- **Mixup**: mixup is a simple and effective data augmentation method and is easy to be implemented [11]. In audio processing, any proportion is used to mix two different audios and corresponding labels, which can reduce the influence of noise sample on the model while expanding the data set. In our training, we adopt a mix of alpha value of 0.4. Referring to the method in [14], the mix images are completely trained in the early stage of the training and reduce to half in the later stage of training.
- **SpecAugment**: SpecAugment is a data augmentation method that acts on Mel spectrogram rather than audio [12]. It modifies the spectrum through the three strategies of distorting the time-domain signal, masking the frequency-domain channel, and masking the time-domain channel. This method can be used to increase the robustness of the network and some fragments in the frequency domain of

[1] Code: https://github.com/Wangkkklll/DSMN_dcase2023

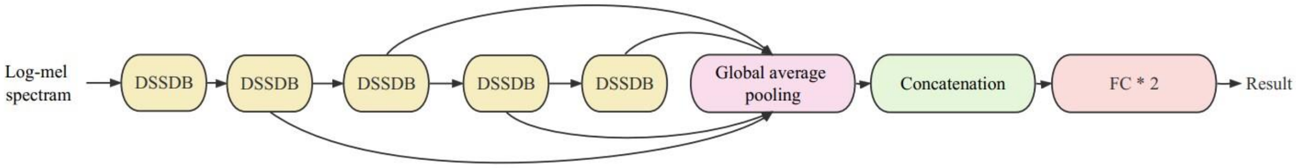

Figure 1: Network Architecture

time and some fragments in the domain. In our training, we apply two masking lines to the time domain and frequency domain. The maximum width of each line is 2.

- **Spectrum correction:** In the datasets provided by the competition, most audios are recorded with device A. Therefore the equipment data is unbalanced. In order to solve this problem, we refer to the method of spectrum correction, which was proposed to solve the problem of mismatching record equipment [13]. We first conduct an average of the spectrum of all devices except device A to obtain a reference device spectrum. Then we use the reference device spectrum to correct the spectrum of device A for additional data. This method is proven to be effective in [13].

3. OUR MODEL

3.1. Network Architecture

The overall structure of our model is shown in Figure 1. We refer to the blueprint separable residual network structure in [15] and improve it based on our ASC task. In order to reduce the complexity of the model, we propose a depth -separated convolution module as the basic composition unit. The input log-Mel spectrogram is deeply processed through the model backbone which is composed of five basic unit cascades.Then the channel splicing method is used to integrate different scale information of the high-level and low-level networks to achieve better performance.

In addition, we have applied multi task learning method. We divide the input audio scenarios into three categories, and train a high-performance three-class classification. Finally, the three classification outputs and the ten classification outputs with higher difficulty are combined to determine the final category.

3.2. Deep Space Separable Distillation Block (DSSDB)

The Figure 2 is our proposed module

- **DSSC:** This structure was used in the DCASE 2020 challenge and achieved better results. Its structure is shown in Figure 3. It can reduce the number of parameters. We do not make a specific introduction here. We use this module for feature enrichment.

- **Conv1:** This is 1 × 1 convolution kernel for dimension transformation. This module can reduce the number of parameters and complete feature distillation.

- **ECA attention:** Wang et al [16] proposed ECA attention, which is an efficient one-dimensional attention mechanism. Compared with the commonly used two-dimensional attention, the amount of parameters and calculations used are relatively small.

For an input feature map, the above-mentioned modules can perform distillation and concentration layer by layer, and finally feature enhancement through attention, which can better extract features. On this basis, combined with the characteristics of the log-Mel spectrum and the characteristics of the distillation block structure, we cut the frequency axis. For the low-frequency part, its characteristics are continuously strengthened, while for the high-frequency part, its features can be concentrated. The overall structure of deep space separable distillation segment block (DSSDB_SEG) is shown in Figure 2.

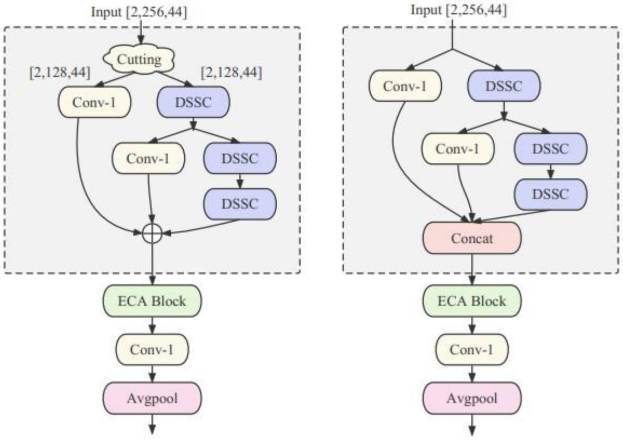

Figure 2: Structure of DSSDB_SEG (left) and DSSDB (right)

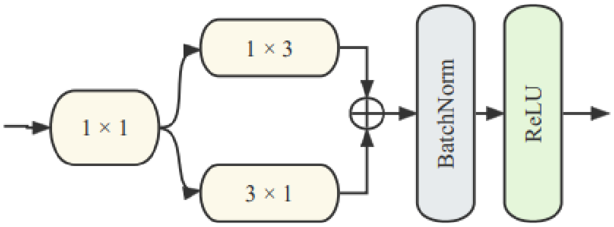

Figure 3: Structure of DSSC

In general, such a structure reduces the number of parameters and MACs a lot. Moreover, since the high and low frequencies are separated, fewer features need to be extracted for each convolution. Hence, it requires fewer weight bits, and the error after the Int8 quantization is smaller at this time.

3.3. Muiti task learning

We added additional labels to each sample in the dataset, and the labels are "Indoor", "Outdoor" and "Traffic". Table 1 shows the mapping relationship between 10 categories and new categories. Therefore, we also need to train a three -class classifier. To reduce memory overhead, we adopted the strategy of model sharing.

One problem of multi -task learning (MLT) is that the gradient conflict may occur in the optimization of different tasks. So we apply the projecting conflicting gradients (PCGrad) method [17]. When there is a problem of gradient conflict between the two tasks, the gradient of one task is projected into the orthogonal direction of the gradient of another problem.

The gradient correction formula is as follows:

$$g_i^{PC} = g_i^{PC} - \frac{g_i^{PC} \cdot g_j}{||g_j||^2} g_j \tag{2}$$

where g_j denotes the gradient of task j, and g_i^{PC} represents the gradient of task i which needs to be corrected.

Table 1: The labels of multi task

	3 new classes	10 classes
1	Indoor	Airport, Metro station, Shopping mall
2	Outdoor	Park, Street, Pedestrian, Public square, Street traffic
3	Transport	Bus, Metro, Tram

The schematic diagram of multi task learning is shown in Figure 4.

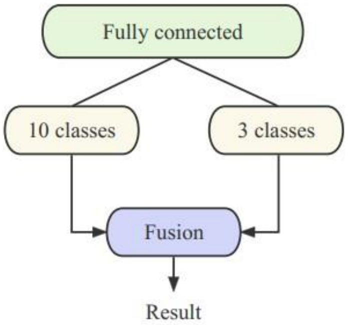

Figure 4: Structure of multi task learning

4. EXPERIMENTAL SETUP

All experiments in this work are conducted using the toolkit of PyTorch. The optimizer is the stochastic gradient descent, and the categorical cross-entropy loss is used. All models are trained 100 epochs with a batch size of 32. In addition, the learning rate is set to 1e-3, using update strategy of Cosine LRScheduler. We use the checkpoint with the highest validation accuracy as the best model.

5. MODEL QUANTIZATION AND EXPERIMENTAL RESULTS

5.1. Model Quantization

According to the DCASE 2023, the limit of space-complexity for the model is 128 KB excluding zero parameters. When using float-point operation with 32 bits, the model contains up to 32768 parameters. This is seriously inconsistent with the parameters of the model we designed. Therefore, in order to ensure the amount of the original parameters, we have adpoted the method of quantitative perception training (QAT) [18]. The method uses float-point operation with 32 bits to train the model in PyTorch, and simulates the quantification effect of Int8 by clamping and spoiling. Because of considering and simulating quantitative errors during the training process, QAT can usually get higher accuracy compared to other quantitative methods.

5.2. Experimental Results

A brief description of the models we submitted is given in Table 2, and the specific models refer to our open source code.

Table 2: Detailed description of the model

	Description
Model1	5*DSSDB_SEG
Model2	5*DSSDB
Model3	1*DSSDB_SEG+4*DSSDB
Model4	5*DSSDB_SEG with big channels

The parameters, MACs, and the accuracy before and after quantization of the model are listed in Table 3.

Table 3: Experimental results of the experiment

	Parms	MACs	Acc (Float32)	Acc (Int8)
Model1	45.1k	8.64M	53.4	53.3
Model2	56.1k	16.74M	59.2	56.4
Model3	56.5k	25.44M	53.4	50.8
Model4	121.8k	20.92M	54.7	52.4

It can be seen from the experimental results that the loss of precision after quantization is smaller when using the frequency-segmented model.

For the submitted model 1, its confusion matrix on the validation set is shown in Figure 5. It can be seen from the confusion matrix that model 1 has a relatively high prediction accuracy for samples in the category of parks.

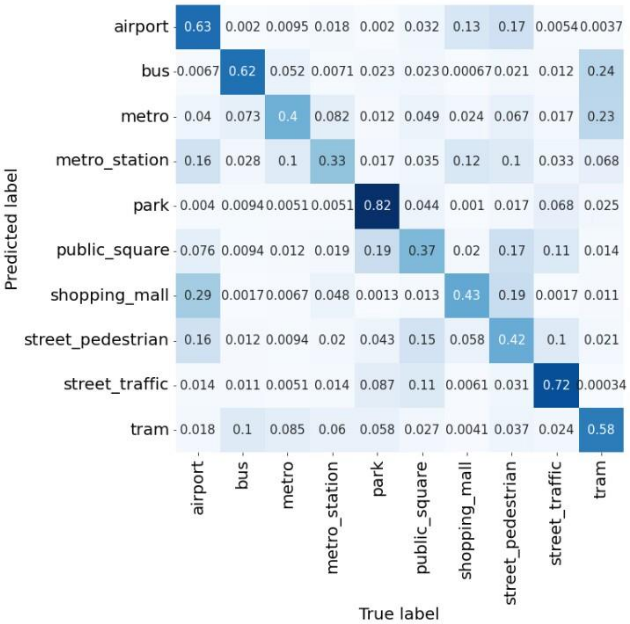

Figure 5: Confusion matrix for the submitted model 1

The logloss scores of the four submitted models in each type of device are listed in Table 4.

Table 4: Logloss scores of the submitted models

Device	Model1	Model2	Model3	Model4
A	1.060	0.974	1.022	1.100
B	1.400	1.283	1.354	1.449
C	1.257	1.135	1.234	1.365
S1	1.299	1.166	1.487	1.383
S2	1.247	1.104	1.310	1.317
S3	1.253	1.053	1.349	1.344
S4	1.373	1.327	1.865	1.476
S5	1.326	1.292	1.894	1.403
S6	1.368	1.389	1.776	1.474

6. CONCLUSIONS

In this technical report, we described a system for the low-complexity acoustic scene classification Task 1 of DCASE challenge 2023. We proposed a deep space separable distillation block (DSSDB) and deep space separable distillation block segment (DSSDB_SEG) as the basic unit. In our network architecture, we applied multi-scale embedding and multi-task learning method to improve the performance. And quantitative perception training (QAT) was used to reduce model complexity. Our experiments showed that using DSSDB as the basic unit can achieve better precision before quantization, while using DSSDB_SEG can maintain a small loss of precision after quantization. Finally, our model achieved a log-loss of 1.287 and an accuracy of 53.3% with the 45.1 KB model size.

7. REFERENCES

[1] I. Martín-Morató, F. Paissan, A. Ancilotto, T. Heittola, A. Mesaros, E. Farella, A. Brutti, and T. Virtanen. Low-complexity acoustic scene classification in dcase 2022 challenge. 2022.

[2] Y. Li, X. Li, Y. Zhang, W. Wang, M. Liu, X. Feng, Acoustic scene classification using deep audio feature and BLSTM network, Proc. ICALIP, July 16-17, 2018, Shanghai, China, 2018, pp. 371-374.

[3] Y. Li, M. Liu, W. Wang, Y. Zhang, Q. He, Acoustic scene clustering using joint optimization of deep embedding learning and clustering iteration, IEEE Transactions on Multimedia, vol. 22, no. 6, pp. 1385-1394, Jun. 2020.

[4] H. K. Chon, Y. Li, W. Cao, Q. Huang, W. Xie, W. Pang, J. Wang, Acoustic scene classification using aggregation of two-scale deep embeddings, IEEE 21st International Conference on Communication Technology, Tianjin, China, Oct. 13-16, 2021, vol. 4, pp. 1341-1345.

[5] Y. Li, W. Cao, W. Xie, Q. Huang, W. Pang, Q. He, Low-complexity acoustic scene classification using data augmentation and lightweight ResNet, in Proc. of The 16th IEEE International Conference on Signal Processing, Beijing, China, Oct. 21-24, 2022, pp. 41-45.

[6] W. Xie, Q. He, Z. Yu, Y. Li, Deep mutual attention network for acoustic scene classification, Digital Signal Processing, vol. 123, 103450, pp. 1-13, April 30, 2022.

[7] R.Sugahara, R. Sato, M. Osawa, Y. Yuno, C. Haruta, (2022). Self-Ensemble with Multi-Task Learning for Low-Complexity Acoustic Scene Classification [Techreport]. DCASE2022 Challenge.

[8] T. Heittola, A. Mesaros, and T. 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), 56–60. 2020.

[9] G. Puy, H. Jain, A. Bursuc. Separable Convolutions and Test-Time Augmentations for Low-Complexity and Calibrated Acoustic Scene Classification [Techreport]. DCASE2021 Challenge, 2021.

[10] Z. Li, Y. Liu, X. Chen, H. Cai, J. Gu, Y. Qiao, C. Dong,. Blueprint Separable Residual Network for Efficient Image Super-Resolution, 2022.

[11] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” arXiv preprint arXiv:1710.09412, 2017.

[12] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: a simple data augmentation method for automatic speech recognition,” in Proc.ISCA Interspeech, 2019, pp. 2019-2680.

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

[14] H. Yu, H. Wang, Jianxin Wu, “Mixup Without Hesitation, ” arXiv:2101.04342. 2021.

[15] Z. Li et al., "Blueprint Separable Residual Network for Efficient Image Super-Resolution," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 2022, pp. 832-842, doi: 10.1109/CVPRW56347.2022.00099.

[16] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 11531-11539, doi: 10.1109/CVPR42600.2020.01155.

[17] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman and C. Finn, " Gradient Surgery for Multi-Task Learning, " arXiv:2001.06782, 2020.

[18] A. D. Kozlov, I. A. Lazarevich, V. Shamporov, N. Lyalyushkin,and Y. Gorbachev, "Neural network compression framework for fast model inference," ArXiv, vol. abs/2002.08679, 2020.