DATA-EFFICIENT ACOUSTIC SCENE CLASSIFICATION VIA ENSEMBLE TEACHERS DISTILLATION AND PRUNING

Technical Report

Shuwei Zhang[1], *Bing Han*[2], *Anbai Jiang*[3], *Xinhu Zheng*[2]
Wei-Qiang Zhang[3], *Xie Chen*[2], *Pingyi Fan*[3], *Cheng Lu*[4], *Jia Liu*[1,3], *Yanmin Qian*[2]

[1] Huakong AI Plus Company Limited, Beijing, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] Tsinghua University, Beijing, China
[4] North China Electric Power University, Beijing, China
eophoeny@gmail.com

ABSTRACT

The goal of the acoustic scene classification task is to classify recordings into one of the ten predefined acoustic scene classes. In this report, we describe the submission of the THU-SJTU team for Task 1 Data-Efficient Low-Complexity Acoustic Scene Classification of the DCASE 2025 challenge. Our methods are consistent with those of last year. Firstly, we use an architecture named SSCP-Mobile (spatially separable), which enhances the CP-Mobile with spatially separable convolution structure and achieves lower computation expenses and better performance. Then we adopt several pre-trained PaSST models as ensemble teachers to teach CP-Mobile with knowledge distillation. After that, we use model pruning techniques to trim the model to meet the computational and parameter requirements of the competition. Finally, we will use knowledge distillation techniques again to fine-tune the pruned model and further improve its performance. Due to some reasons, our submissions included four systems that contain only general models, but we also attempted to use device type information to increase the performance of the system S1.

Index Terms— SSCP-Mobile, ensemble teachers, model pruning, acoustic scene classification, low-complexity

1. INTRODUCTION

The task of acoustic scene classification (ASC) is to classify recordings into one of the ten predefined acoustic scene classes. In task 1 of DCASE2025 [1], participants must design a low-complexity network that can predict scenes for one second of audio. This year, the organizer encourages participants to use data-efficient approaches and exploit device type information.

In task 1 of DCASE2025, the key challenges of this task are still the “data-efficient” and “low-complexity” problems, and data efficiency is gaining more focus in 2025. The main differences in the task for this year are as follows.

- Device information is available for evaluation, allowing participants to fine-tune models for specific recording devices.
- Models must be trained only on the 25% subset, encouraging data-efficient approaches.
- No restrictions on external datasets.
- Participants are required to submit inference code.

In order to achieve higher accuracy with limited parameter and computational limitations, we mainly develop our challenge system from the aspects of model design, training strategies, model pruning, and distillation. Although device-specific models are not contained in our submissions, fine-tuning based on device information can improve the performance of the general model. In the following sections, we will provide a brief description of our challenge systems.

2. APPROACHES

We will introduce the system description from the following aspects:

2.1. Model Architecture

Our baseline student architecture is the SSCP-Mobile model, which is based on the low-complexity CP-Mobile model described in [2]. The most expensive operations in CP-Mobile are 3x3 convolution operations. In SSCP-Mobile, each 3x3 convolution is replaced by a fusion of 1x3 and 3x1 convolution layers.

Table 1: Configuration of three submission systems.

Sys ID	Para. Num	MMACs	Teacher Models	Pruned Metric	Pruned Others
S1	63748	29982132	4 PaSST	AGP	Pruned from base channel 64 to 32
S2	63748	29982132	4 PaSST	AGP	Pruned from base channel 64 to 32
S3	63215	29221122	4 PaSST	AGP	Progressive pruned from base channel 96 to 64 then to 32
S4	63215	29221122	4 PaSST	AGP	Progressive pruned from base channel 96 to 64 then to 32

Table 2: Performance test of four submission systems. Systems only contain general models.

Sys ID	Overall	a	b	c	s1	s2	s3	s4	s5	s6
S1	**59.04**	68.81	**60.58**	63.92	**57.30**	55.12	**60.58**	57.42	**55.94**	51.70
S2	58.54	69.81	60.55	64.71	55.03	55.63	59.76	56.61	54.81	49.94
S3	58.95	69.70	59.48	**64.77**	57.12	**56.67**	59.97	57.30	55.27	50.30
S4	58.73	**70.09**	59.64	64.44	54.58	55.36	59.54	**57.63**	55.30	**52.00**

2.2. Data Preprocess

2.2.1. Feature extraction

For CNN based student model SSCP-Mobile, following the baseline setup in[1], we use audio at a sampling rate of 32 kHz to compute Mel-Spectrograms with 256 frequency bins. Short Time Fourier Transformation (STFT) is applied with a window size of 96 ms and a hop size of 16 ms.

For transformer [3] based teacher model PaSST [4], in order to utilize the pre-trained model parameters[2], we use the same feature configuration as the original paper.

2.2.2. Data Augmentation

Data augmentation has a crucial impact on model performance, and we have adopted the following data augmentation strategies.

- Roll Audios: In order to enhance the diversity of training data, we roll audio clips to achieve better performance.
- SpecAug [5]: We will apply random masking in the frequency band and time dimensions to enhance the robustness and generalization of the model.
- Freq-MixStyle [6, 2]: Freq-MixStyle (FMS) is a frequency-wise version of the original MixStyle [7] that operates on the channel dimension. FMS normalizes the frequency bands in a spectrogram and then denormalizes them with mixed frequency statistics of two spectrograms. FMS is applied to a batch with a certain probability specified by the hyperparameter p_{FMS} and the mixing coefficient is drawn from a Beta distribution parameterized by a hyperparameter α.

2.3. Ensemble Teachers for Knowledge Distillation

To improve the performance of the small model, we adopted distillation with the teacher-student paradigm. We chose PaSST to teach convolutional-based baseline models SSCP-Mobile. Four PaSST models are selected for knowledge distillation.

2.4. Model Pruning

In addition, to build more efficient low-complexity models, we also adopted pruning strategies to reduce the complexity of the model. Our overall process will be divided into multiple steps.

- Firstly, based on model distillation, we construct multiple relatively larger SSCP-Mobile models (such as the base channels are 48, 64, 80, 96, respectively).
- Then, we use model pruning strategy to prune larger models into standard models whose computational and parameter requirements meet the requirements of the challenge.
- Finally, we use knowledge distillation to fine-tune the pruned model for further improvements.

2.5. Device-Specific Fine-Tuning

Device-specific models are not contained in our submissions, but fine-tuning based on device information can increase the performance of the general model. Both hard labels and soft labels derived from knowledge distillation can be utilized for fine-tuning.

[1] `https://github.com/marmoi/dcase2022_task1_baseline`

[2] `https://github.com/kkoutini/PaSST`

Table 3: Comparative experiments on the general model and the device-specific models of system S1.

Device-id	a	b	c	s1	s2	s3
General	68.81	60.58	63.92	57.30	55.12	60.58
Device-Specific	71.21	63.06	66.78	59.27	56.81	61.21
Improvement	2.40	2.48	2.86	1.97	1.69	0.63

3. SUBMISSIONS AND RESULTS

The configuration differences of the four systems that we submitted are shown in Table 1, and the performance is presented in Table 2. Table 3 illustrates the performance gains for system S1 achieved by device-specific fine-tuning.

4. REFERENCES

[1] http://dcase.community/challenge2025/.

[2] F. Schmid, T. Morocutti, S. Masoudian, K. Koutini, and G. Widmer, "Cp-jku submission to dcase23: Efficient acoustic scene classification with cp-mobile," DCASE2023 Challenge, Tech. Rep, Tech. Rep., 2023.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] K. Koutini, J. Schlüter, H. Eghbal-Zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," *arXiv preprint arXiv:2110.05069*, 2021.

[5] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[6] B. Kim, S. Yang, J. Kim, H. Park, J. Lee, and S. Chang, "Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification," *arXiv preprint arXiv:2206.12513*, 2022.

[7] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," *arXiv preprint arXiv:2104.02008*, 2021.