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

Technical Report

Bing Han[1]*, Wen Huang*[1]*, Zhengyang Chen*[1]*, Anbai Jiang*[2]*, Xie Chen*[1]*, Pingyi Fan*[2]*,*
Cheng Lu[3], Zhiqiang Lv[4], Jia Liu[2,4] Wei-Qiang Zhang[2], Yanmin Qian[1]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Tsinghua University, Beijing, China
[3] North China Electric Power University, Beijing, China
[4] Huakong AI Plus Company Limited, Beijing, China
{hanbing97, yanminqian}@sjtu.edu.cn

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 SJTU-THU team's submission for Task 1 Data-Efficient Low-Complexity Acoustic Scene Classification of the DCASE 2024 challenge. Firstly, we design a new architecture named SSCP-Mobile (spatially separable) by enhancing the CP-Mobile with spatially separable convolution structure and achieve lower computation expenses and better performance. Then we adopt several pretrained 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.

***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 DCASE2024 [1], it requires participants to design a low-complexity network that can predict scenes for one second of audio. Different from the previous editions of the challenge focused on two important problems: (1) recording device mismatch and (2) low-complexity constraints. This year, the organizer hopes to tackle an additional challenging real-world situation: the limited availability of labeled data. They encourage participants to design systems that maintain high prediction accuracy while using as little labeled data as possible.

In task 1 of DCASE2024, the key challenges of this task are the "data-efficient" and "low-complexity" problems. In practical scenarios, it is difficult for us to collect data and train models on a new machine, or the number of machines is very small. So the main differences in the task of this year are that:

- Training sets of different sizes are provided. These train subsets contain approximately 5%, 10%, 25%, 50%, and 100% of the audio snippets in the train split provided in Task 1 of the DCASE 2023 challenge. A system must only be trained on the specified subset and the explicitly allowed external resources.
- There exist hard complexity limits in terms of model size (128 kB) and multiply accumulate operations (30 Million MACs)

In order to achieve the highest possible accuracy with limited parameter and computational limitations, we mainly design our challenge system from the aspects of model design, training strategies, model pruning, and distillation. 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 Design

Our baseline student architecture is based on the low-complexity CP-Mobile model described in [2]. We redesign the model to increase its representation capability and efficiency with spatially separable convolution structure and call the final model SSCP-Mobile (CPM). The most expensive

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	63875	29840890	4 PaSST	Linear	Pruned from base channel 96 to 32
S3	63215	29221122	4 PaSST	Agp	Progressive pruned from base channel 96 to 64 then to 32

Table 2: Performance evaluation of three submission systems on different split data.

Sys ID	5	10	25	50	100
S1	0.521763	0.558267	0.588725	**0.614157**	**0.621780**
S2	**0.529521**	**0.569413**	**0.591428**	0.609491	0.613431
S3	0.527011	0.558442	0.587392	0.603626	0.617386

operations in CP-Mobile are 3x3 convolutions operation. We replace each 3x3 convolution with a fusion of 1x3 and 3x1 convolution layers.

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.

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

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

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. Among the pre-training models provided by the official, we selected three models that were finetuned on the training set as teacher models. In addition, in order to diversify teachers, we have also retrained a Passt from scratch as one of the teachers. Finally, we ensemble these four models for knowledge distillation.

We also tried BEATs [8] and CP-Mobile, but these two models performed poorly when there was limited training data.

2.4. Model Pruning

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

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

In addition, we also attempted progressive pruning, cutting from large models to relatively large models and then cutting to sizes that meet the requirements.

3. SUBMISSIONS AND RESULTS

The configuration differences of the three systems we submitted are shown in Table 1, and the performance is presented in Table 2.

4. REFERENCES

[1] http://dcase.community/challenge2024/.

[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.

[8] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.