EFFICIENT ACOUSTIC SCENE CLASSIFICATION USING MEAN-TEACHER AND KNOWLEDGE DISTILLATION

Technical Report

Maxim K. Surkov

ITMO University, Saint Petersburg, Russia

ABSTRACT

This technical report describes submission for the DCASE 2024 Task 1 "Data-Efficient Low-Complexity Acoustic Scene Classification". Pretrained Efficient-AT was fine-tuned using a mean-teacher self supervised learning algorithm on labeled data, presented by the authors of the task, and on external unlabeled data taken from AudioSet. Current state-of-the-art efficient neural network CP-Mobile was then trained on the same data using knowledge distillation from fine-tuned Efficient-AT model. Proposed model consists of 61K parameters and requires 22M MACs. Using sufficient amount of external data in pair with knowledge distillation improve the results by around 4% in accuracy in compare with the baseline approach in cases with the small amount of labeled data.

***Index Terms*—** Mean-teacher, knowledge distillation, Efficient-AT, CP-Mobile

1. INTRODUCTION

In DCASE 2024 Task 1 "Data-Efficient Low-Complexity Acoustic Scene Classification" [1] authors presented a problem of designing a system with a restricted amount of memory, MACs and labeled data. Model size is bounded by 128Kb, MACs should not be greater than 30M operations. A new restriction has been added this year. There exist 5 cases that differ from each other in the amount of labeled data: 5%, 10%, 25%, 50% and 100% of the total training dataset. Baseline system is the CP-Mobile [2], fine-tuned on the presented labeled data [3]. This report presents an approach based on the self-supervised learning algorithm called mean-teacher [4] which allows the use external unlabeled data in order to improve the accuracy of the model's predictions. Training process maintains an exponential average of the model which is called teacher-model and used for self distillation to the student-model. External subsets of 10k and 20k from AudioSet were extracted using a specifically designed set of labels. Unlabeled data was used together with labeled data, presented by the authors of the task, in order to fine-tune Efficient-AT [5], pretrained on AudioSet. Finally, CP-Mobile was trained using knowledge distillation from the fine-tuned Efficient-AT on both labeled and unlabeled dataset parts.

2. DATA PREPROCESSING

In all experiments, a sampling rate of 16kHz was used to compute Mel spectrograms with 128 frequency bins. Short Time Fourier Transformation (STFT) is applied with a window size 128 ms and a hop size of 10 ms. Spectrogram augmentations mask 15% of the input Mel spectrogram along both the time and frequency dimensions. MixStyle augmentation [6] was applied with the following parameters: $p = 0.4, \alpha = 0.3, \varepsilon = 10^{-6}$.

In order to extend the number of training data samples, two random subsets of AudioSet with the sizes of 10K (small) and 20K (large) samples were extracted using labels presented in Table 1.

code	label
/m/012f08	Motor vehicle (road)
/m/014yck	Aircraft engine
/m/0195fx	Subway, metro, underground
/m/01bjv	Bus
/m/01g50p	Railroad car, train wagon
/m/06d_3	Rail transport
/m/07jdr	Train
/m/07yv9	Vehicle
/m/09ddx	Ducks, geese, waterfowl
/m/0btp2	Traffic noise, roadway noise
/m/0cmf2	Fixed-wing aircraft, airplane
/t/dd00092	Wind noise (microphone)
/t/dd00127	Inside, public space
/t/dd00128	Outside, urban or manmade

Table 1: Audioset labels used to extract unlabeled subsets

Input (freq×time×channels)	Operator	Stride
$128 \times 101 \times 1$	Conv2D 3×3, BN, ReLU	2×2
$64 \times 51 \times 8$	Conv2D 3×3, BN, ReLU	2×2
$32 \times 26 \times 32$	CPM Block S	1×1
$32 \times 26 \times 32$	CPM Block S	1×1
$32 \times 26 \times 32$	CPM Block D	2×1
$16 \times 26 \times 32$	CPM Block T	1×2
$16 \times 13 \times 56$	CPM Block S	1×1
$16 \times 13 \times 56$	CPM Block T	1×1
$16 \times 13 \times 56$	Conv2D 1×1, BN, Avg. Pool	

Table 2: CP-Mobile architecture

3. MODELS

Efficient-AT was used with the basic configuration called *mn10_as* as a teacher model. CP-Mobile with the basic configuration, described in Table 2 was used as a student model.

4. EXPERIMENTS

Model	Split 5%	Split 10%	Split 25%	Split 50%	Split 100%	Params	MACs
Baseline	42.40	45.29	50.29	53.19	56.99	61.148K	29.419M
Efficient-AT	43.49	47.26	51.86	55.45	57.99	4.214M	62.722M
distilled CP-Mobile (small subset)	**46.47**	48.95	51.77	53.96	55.56	61.148K	21.896M
distilled CP-Mobile (large subset)	46.17	**49.61**	**53.5**	**54.68**	**55.82**	61.148K	21.896M

Table 3: Experimental results

Whole training pipeline consists of 2 stages. Firslty, the pretrained Efficient-AT model was fine-tuned on a labeled dataset together with the unlabeled AudioSet subset consisting of 10K examples using self-supervised mean-teacher algorithm. Exponential mean average factor was set to the default value of 0.999. During mean-teacher, linear combination of supervised and self-supervised losses was used (see Equation 1). Exponential ramp up scheduler was chosen with the following parameters: 10% of exponential ramp up steps, 10% of constant weight and then α linearly decays to zero.

$$L = L_{sup} + \alpha \cdot L_{self\_sup} \tag{1}$$

Secondly, CP-Mobile was fine-tuned on labeled data together with both small and large Audioset subsets using knowledge distillation. Mean square error between teacher and student predicted probabilities was selected as the loss function to optimize. Learning rate scheduler is the cosine annealing with linear warmup during the first 10% of training steps, maximum learning rate is equal to 10^{-3}. Model weights were optimized with the AdamW optimizer with the weight decay of 10^{-3}. All models were trained during 750 epochs with batch size of 256. Experimental results can be found in Table 3.

5. CONCLUSION

In this technical report, two stage training pipeline was used to compensate the lack of labeled data. Firstly, pretrained Efficient-AT was fine-tuned on both labeled data, presented by authors and external data, retrieved from the AudioSet using mean-teacher self supervised learning algorithm. Secondly, Efficient CP-Mobile was fine-tuned on large amount of unlabeled data with the knowledge distillation approach. Overall, obtained system can classify acoustic scenes with high quality in low-data cases.

6. REFERENCES

[1] F. Schmid, P. Primus, T. Heittola, A. Mesaros, I. Martín-Morató, K. Koutini, and G. Widmer, "Data-efficient low-complexity acoustic scene classification in the dcase 2024 challenge," 2024. [Online]. Available: https://arxiv.org/abs/1706.10006

[2] F. Schmid, T. Morocutti, S. Masoudian, K. Koutini, and G. Widmer, "Distilling the knowledge of transformers and CNNs with CP-mobile," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, 2023, pp. 161–165.

[3] 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)*, 2020, pp. 56–60. [Online]. Available: https://arxiv.org/abs/2005.14623

[4] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[5] F. Schmid, K. Koutini, and G. Widmer, "Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

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