UPB-NT SUBMISSION TO DCASE24: DATASET PRUNING FOR TARGETED KNOWLEDGE DISTILLATION

Technical Report

Alexander Werning

Paderborn University
Department of Communications Engineering,
Paderborn, Germany
werning@nt.upb.de

Reinhold Haeb-Umbach

Paderborn University
Department of Communications Engineering,
Paderborn, Germany
haeb@nt.upb.de

ABSTRACT

In this technical report, we describe our submission for Task 1 *Data-Efficient Low-Complexity Acoustic Scene Classification* [1]. We adopt the baseline model and add a specialised knowledge distillation process before proceeding with the baseline training process. Our model was distilled on a pruned subset of the AudioSet dataset using large pretrained models. The pruning of the dataset is based on the similarity of the data to the targeted challenge dataset.

Index Terms— data pruning, knowledge distillation

1. INTRODUCTION

The motivation for our approach is the observation that typical pretraining data, such as the AudioSet dataset [2] contain a wide range of different acoustic event types, much more than are required for a specific task. If the capacity of a neural network for a target application is highly constrained, it may be beneficial to only use relevant data for its training to avoid spending modeling capacity for irrelevant event types and instead train on a subset that is selected to match the targeted application. Since datasets such as AudioSet are very large, a selected matching subset can still be much larger than the data that is available for a specific task. It was shown that for a dataset of environmental sounds, that distilling on a portion (5%) of AudioSet produced better results than using the whole dataset [3]. We apply the approach presented in [3] to the challenge data [4] of Task 1 to prune AudioSet and apply knowledge distillation on this subset.

2. METHOD

Our method consists of three steps. First, a domain classifier is learned which is used to find a relevant portion of the data in AudioSet. Then, knowledge distillation using an ensemble of teacher models is performed. Finally, the student model is fine-tuned to the challenge dataset.

2.1. Data Domain Classifier

Based on embeddings of the AudioSet and the challenge datasets, a domain classifier is learned. Embeddings of the audio examples are computed as the logits of a pre-trained nine-model ensemble of PaSST audio tagging models [5]. The domain classifier, for which we choose a linear classifier, operates on these embeddings. The model is trained using Adam. The training data consists of the evaluation portion (*eval*) of AudioSet and the respective training subset of the challenge data. The AudioSet data is labeled as negative, or not matching the challenge data (0), while the challenge data is labeled positive (1). Each minibatch of training data is sampled such that it contains an equal amount of positive and negative samples. The batch size is set to 600, and the training duration is 400 epochs. The learning rate is set to 10^{-4}. The decision how much AudioSet data to keep is determined by a threshold value.

2.2. Knowledge Distillation

The approach is based on the description in [3]. Knowledge distillation is performed using a PaSST model ensemble [5] as a teacher and the baseline CP-Mobile model [6] as the student model. The last layer of the model is adapted to the number of classes of AudioSet. The distillation is performed offline, that is the logits of the teacher are pre-computed once and later loaded from disk. We apply both a distillation loss and a classification loss using the provided AudioSet labels. A higher weighting factor $\lambda = 0.9$ is applied to the distillation loss. A batch size of 32 is used for 100 k training iterations. The maximum learning rate is 5×10^{-3} which then follows a learning rate schedule of a cosine decay to 1% of the maximum learning rate.

The sampling rate of the audio data is 32 kHz, the models are trained on mel spectrogram features with 256 mel bins, window size of 3072 samples, and a hop size of 500 samples. The data is augmented using MixUp [7] and SpecAugment [8], this only affects the input to the student as the teacher outputs are pre-computed.

2.3. Fine-tuning

The student model from the knowledge distillation is fine-tuned to the challenge data. The last layer is adapted and re-initialized randomly to match the number of classes. For the actual model training, the settings of the baseline are used as described in [1].

The SpecAugment and Frequency MixStyle augmentations from the baseline are kept, just the duration of the training is shortened to 100 epochs.

3. RESULTS

Results for different thresholds on the full training split are shown in Table 1. The threshold value used for the dataset pruning was then chosen as 0.3 for later experiments.

Table 1: Results of the model on the test set of the challenge dataset for different pruning threshold values on the full AudioSet.

Threshold	Macro Avg Acc (%)	Relative size of pruned set
0.1	57.8	30.9%
0.2	56.7	19.2%
0.3	60.0	14.2%
0.4	59.1	10.9%
0.5	58.1	8.5%

The challenge task requires an evaluation using five splits of the training data (5, 10, 25, 50, 100). The data pruning of our approach is dependent on the training data, so we perform five independent data pruning operations and subsequent distillations. The results of the distilled and fine-tuned models is shown in Table 2.

Table 2: Results of the model on the test set of the challenge dataset.

Split	Macro Avg Acc (%)	Baseline
5	45.8	42.4
10	50.3	45.3
25	53.8	50.3
50	55.0	53.2
100	60.0	57.0

4. CONCLUSION

In this report we described the UPB-NT submission to the DCASE Challenge 2024 Task 1. We performed a knowledge distillation on a pruned subset of the AudioSet database using ensembles of pre-trained PaSST models to obtained a pre-trained model specialised to the given task. This model was further fine-tuned using the challenge data.

5. ACKNOWLEDGMENT

The project was funded by the Federal Ministry of Education and Research (BMBF) under grant no. 01IS22094E WEST-AI.

The authors gratefully acknowledge the computing time provided to them on the high-performance computers Noctua 1/2 at the NHR Center PC2. These are funded by the German Federal Ministry of Education and Research and the state governments participating on the basis of the resolutions of the GWK for the national high-performance computing at universities (www.nhr-verein.de/unsere-partner)

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] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[3] A. Werning and R. Haeb-Umbach, "Target-specific dataset pruning for compression of audio tagging models," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2024, accepted at EUSIPCO 2024.

[4] 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

[5] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient Training of Audio Transformers with Patchout," in *Interspeech, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 2753–2757. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-227

[6] 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.

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

[8] 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 *Interspeech, 20th Annual Conference of the International Speech Communication Association, Graz, Austria*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 2613–2617. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-2680