SUBMISSION TO DCASE 2023 TASK 1: DEVICE INVARIANT TRAINING WITH STRUCTURED FILTER PRUNING FOR LOW COMPLEXITY ACOUSTIC SCENE CLASSIFICATION

Technical Report

Lorenz P. Schmidt, Beran Kiliç, Nils Peters

Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
International Audio Laboratories, Erlangen, Germany
{lopa.schmidt, beran.kilic, nils.peters}@fau.de

ABSTRACT

This technical reports describes our contribution to the DCASE challenge 2023 Acoustic Scene Classification Task 1. We apply Inverse Contrastive Learning to regularize models and generalize better to unseen devices. First we construct a teacher ensemble by fine-tuning several PaSST models and then train student models with different Memory-Accumulate Counts (MACs) hard constraints. This yields four different models with approximately MMACs of 30, 20, 10 and 5. Finally the model is quantized to 8bit in order to fulfill memory requirements of the challenge.

Index Terms— Contrastive Learning, Alternating Direction Method of Multipliers (ADMM), Patchout Audio Transformer (PaSST), Receptive Field Regularization

1. INTRODUCTION

The DCASE Challenge for Acoustic Scene Classification poses a yearly task with different difficulties. In 2022 the challenge difficulty increases significantly by reducing the sample length from 10s to only 1s. This year changed leaderboard assessment and takes not only the performance in consideration, but also the ranking for number of parameters and Multiplication-Accumulate Counts (MACs).

In order to prune and quantize models at different hard constraints of MMACs and parameter counts, we apply constrain proximal operator with Alternating Direction Method of Multipliers (ADMM) in a student-teacher framework. Further we regularize our teacher model with Inverse Contrastive Learning [1] for the device classes.

We use the model architecture of last year winning team and re-use also their hyperparameter, if not specified otherwise. In the next sections we describe our training procedure.

2. TEACHER MODEL TRAINING

2.1. Optimization Hyperparameter Search

Inspired by the a recent paper [2] indicating the importance of the ϵ parameter when training models with the ADAM optimizer, we first conduct a hyperparameter search for all parameters $(\lambda, \beta_1, \beta_2, \epsilon) \in ((1e-5, 1e-1), (0,1), (0,1), (1e-10, 1e3))$.

The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer IIS.

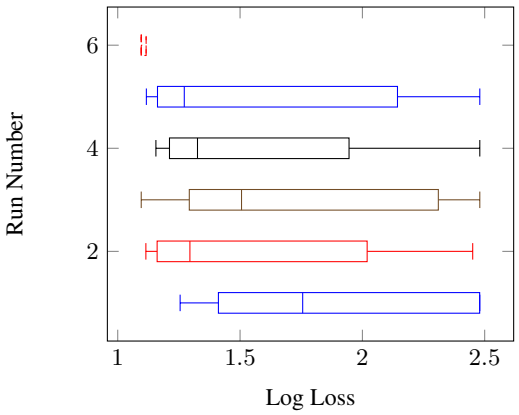

Figure 1: Results of the hyperparameter search for Adam optimizer. In each run 10 different instances are submitted.

2.2. Device Invariant Training

The DCASE challenge task 1 dataset is heavily biased towards a single device (device A), having 73% of the total training data. This induces an implicit bias for the Acoustic Scene Classification task and maximizes reliance on device specific features when minimizing the overall cross entropy.

We take the stance that information on the device class at the final latent (before classification to target labels) is undesired. To make the model invariant to device classes, we penalize latents following distances measures similar to the device classes.

$$\mathcal{L}_{\text{icl}} = \mathop{\mathbb{E}}_{\substack{(z,c)\sim p(z,c) \\ (\hat{z},\hat{c})\sim p(\hat{z},\hat{c})}} [\mathbb{1}(c=\hat{c}) f(z,\hat{z}) + \mathbb{1}(c\neq\hat{c}) g(z,\hat{z})] \quad (1)$$

with a squared distance function $g(z,\hat{z}) = d_Z^2(z,\hat{z})$ pulling same classes, exponential $f(z,\hat{z}) = \exp(\alpha - d_Z(z,\hat{z}))$ pushing latent aways from different classes.

The contrastive loss is applied in tandem to the cross-entropy loss

$$\mathcal{L}(z) = \mathcal{L}_{\text{CE}}(z) + \lambda \mathcal{L}_{\text{ICL}}(z) \quad (2)$$

with the barrier value α and loss weighting factor λ.

α	LogLoss	Accuracy	λ	LogLoss	Accuracy
0.0	1.265	0.5172	0.1	1.139	0.5687
0.2	1.27	0.5128	0.5	1.1	0.5832
0.4	1.313	0.5034	1.0	1.113	0.5852
0.8	1.331	0.4902	3.0	1.139	0.5687
1.5	2.166	0.216	6.0	1.183	0.563

Figure 2: Results for different α and λ values. Based on the results we choose $\alpha = 0.2$, $\lambda = 0.5$

Device	**Ensemble Contrastive**		**W/o Contrastive**	
	LogLoss	Accuracy[%]	LogLoss	Accuracy[%]
A	0.9675	73.06	0.995	72.39
B	1.1630	63.77	1.152	67.25
C	1.0600	71.58	1.078	67.21
S1	1.213	65.99	1.362	58.59
S2	1.441	59.60	1.439	57.91
S3	1.235	62.96	1.311	57.58
S4	1.287	65.32	1.338	62.63
S5	1.278	67.68	1.293	60.27
S6	1.361	61.95	1.453	58.25

Table 1: Generalization of teacher model after training with ICL loss

2.3. Ensemble Learning using Generalized Mean

After finding optimal hyperparameters for the teacher model, we train eight different models with different seeds. To reduce variance of the teacher model further, we calibrate our ensemble with generalized mean:

$$\mathcal{L}_j = \frac{1}{p_j}[\log \sum_i \exp(\mathcal{L}_{ij} p_j + \log w_i) - \log N] \tag{3}$$

The $\log \sum_i \exp(\cdot)$ is expressed with the log-sum-exp operator in PyTorch and the model weight w_i and target exponent p_j learned during training. The models parameters are fixed and not updated in this phase. We get a final result of the ensemble teacher model of 1.085 log-loss and 59.37% accuracy.

3. STUDENT MODEL TRAINING

For the student model we adopt a Receiptive Field Regularized Convolutional Neural Network and is based on the student model of the winning team of last year. In 2023 the DCASE challenge ranks submission based on their total MACs and parameter counts.

Model	Weight	Model	Weight
1	0.0226	5	0.0306
2	0.0224	6	0.1666
3	0.0838	7	0.0322
4	0.0792	8	0.2529

Figure 3: Final weights for eight model ensemble.

3.1. Structured Filter Pruning

The RF-CNN model uses only convolutional layers, except the last layer for classification to the final scene targets. To the end of pruning our model to different MMACs and parameter count, we are separating the cross-entropy loss and non-convex indicator function on the group norm of individual filters

$$\min_{(a,b)} \mathcal{L}(a) + \mathcal{I}_C(b) \quad \text{s.t.} \quad a = b \tag{4}$$

with $\mathcal{I}_C(\cdot)$ the indicator function indicating that the feasible region is restricted to the set $C = \{\theta : |\theta|_{\text{MAC}} \leq A \text{ and } |\theta|_{\text{params}} \leq B\}$. Estimating the optimal thresholding scalar for group norms is non-linear and an algorithm described in Algorithm 1.

The optimization problem is augmented with the projection and gradient ascent of dual parameters.

$$\min_x \tag{5}$$

Algorithm 1 Group filter norm thresholding estimation

Sort group norm $\|\theta_j\|_{\ell_{1,2}}$ with $0 \leq j \leq N$
low, high $\leftarrow 0, N$
while low $\leq$ high **do**
  idx, $c \leftarrow \frac{\text{low+high}}{2}$, $\|\theta_{\text{idx}}\|_{\ell_{1,2}}$
  Mask filter sets where $\|\theta_i\|_{\ell_1 \ell_2} < c_i$
  Evaluate active paths $f(\mathbf{x})$ by setting inputs to NaNs
  if $|\theta|_{\text{MAC}} < A$ and $|\theta|_{\text{param}} <$ B **then**
    high $\leftarrow$ idx $- 1$
  else
    low $\leftarrow$ idx $+ 1$
  end if
end while

3.2. Knowledge Distillation

The final

4. REFERENCES

[1] A. K. Akash, V. S. Lokhande, S. N. Ravi, and V. Singh, "Learning invariant representations using inverse contrastive loss," 2021.

[2] D. Choi, C. J. Shallue, Z. Nado, J. Lee, C. J. Maddison, and G. E. Dahl, "On empirical comparisons of optimizers for deep learning," 2020.