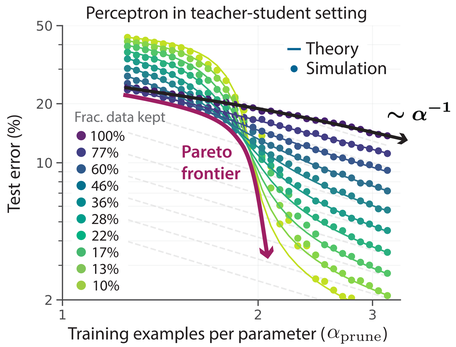
Beyond neural scaling laws: beating power-law scaling via dataset pruning

Ben Sorscher*, Robert Geirhos*, Shashank Shekhar, Surya Ganguli, Ari Morcos

Outstanding paper award!

Summary: By more carefully selecting which examples to train on, we can break through power laws and achieve faster scaling in test loss with dataset size.

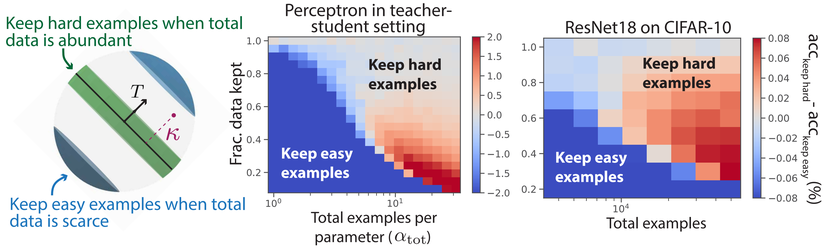


Background:

- *Neural scaling laws*: a growing body of work shows test loss often falls off like a power law with resources like the number of training examples (P) number of parameters (N) or compute (C) [1-7].
- *Data pruning:* recent works show that models can be trained to full performance on a fraction of the training set by ranking examples from "easiest" to "hardest", and pruning away the easy examples [8-11].

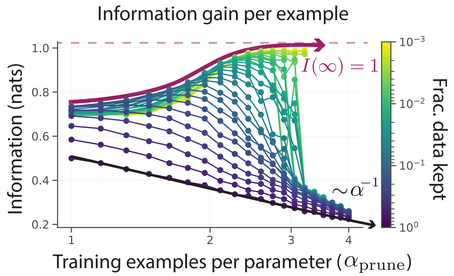
An analytical theory for data pruning: Using replica theory we derive the generalization error of data pruning in the teacher-student perceptron setting, for an arbitrary pruning strategy. This theory yields two key predictions:

1. *The optimal pruning strategy depends on the dataset size.*
2. *Pareto optimal pruning can beat power law scaling and achieve exponential scaling.*

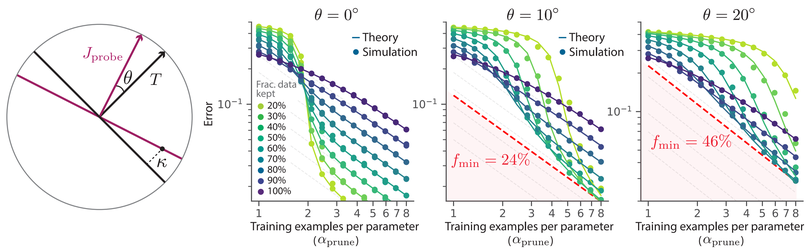


For small datasets, easy examples are crucial to learn coarse features. While for large datasets, hard examples are most informative of the fine-grained structure of the target function.

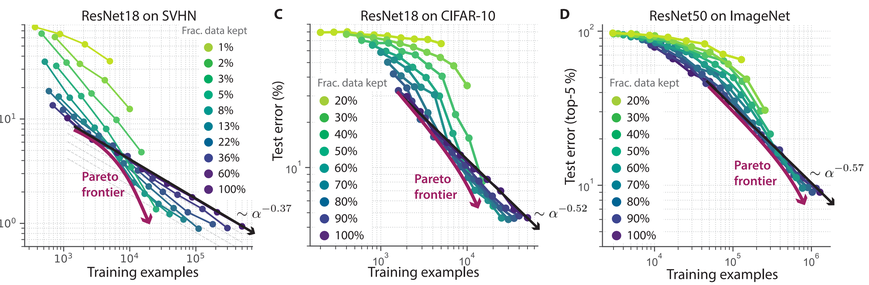
Intuition: when data is drawn i.i.d., the marginal information per example decays to zero as dataset size grows , resulting in power law scaling. Optimal pruning ensures each example provides finite information, leading to exponential decay.



A good pruning metric is key: with an imperfect metric, generalization error will eventually settle onto a power law lower envelope governed by the quality of the pruning metric.

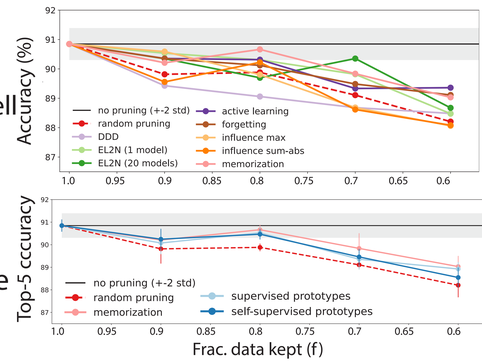


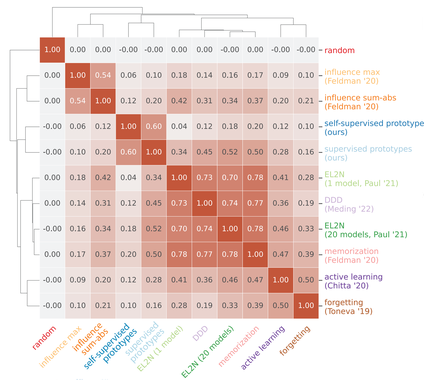
Optimal data pruning breaks scaling laws in practice: we verify the predictions of our theory by performing pareto optimal pruning using ResNets trained on standard vision tasks.

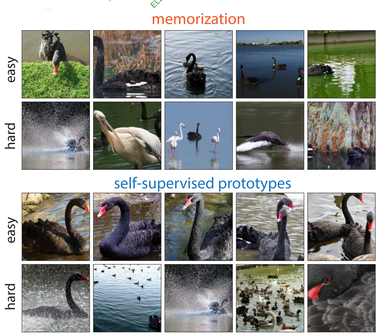


Outlook: how does data pruning perform on massive, uncurated datasets (eg LAION-5B or Common Crawl)?

Data pruning at ImageNet scale: We benchmark a wide range of existing pruning metrics shown to perform well on CIFAR-10 and find mixed results at ImageNet scale. Moreover, existing metrics require labels and are sometimes quite expensive to compute.







A novel self-supervised pruning metric: Motivated by this, we introduce a simple and cheap metric (self-supervised prototypes) for raking training examples. The method is based on k-means clustering in the penultimate layer of a SSL model (SWAV [12]), scoring examples near to a cluster as "easy" and those further away as "hard". It achieves near-original performance when keeping only 80% of ImageNet training data. A question for the future: how much more can we prune ImageNet?

Outlook: towards foundation datasets, where the computational cost of data pruning can be amortized across efficiency gains in training many downstream models.

References: [1] Hestness et al. '17. [2] Kaplan et al. '20. [3] Henighan et al. '20. [4] Gordon et al. '21. [5] Hernandez et al '21. [6] Zhai et al '21. [7] Hoffman et al. '22. [8] Feldman & Zhang '20. [9] Toneva et al. '19. [10] Chitta et al. '21. [11] Paul et al. '21. [12] Caron et al. '20.