Beyond neural scaling laws: beating power law scaling via data pruning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Widely observed neural scaling laws, in which error falls off as a power of the training set size, model size, or both, have driven substantial performance improvements in deep learning. However, these improvements through scaling alone require considerable costs in compute and energy. Here we focus on the scaling of error with dataset size and show how both in theory and practice we can break beyond power law scaling and reduce it to exponential scaling instead if we have access to a high-quality data pruning metric that ranks the order in which training examples should be discarded to achieve any pruned dataset size. We then test this new exponential scaling prediction with pruned dataset size empirically, and indeed observe better than power-law scaling performance on ResNets trained on CIFAR-10, SVHN, and ImageNet. Given the importance of finding high-quality pruning metrics, we perform the first large-scale benchmarking study of ten different data pruning metrics on ImageNet. We find most existing high performing metrics scale poorly to ImageNet, while the best are computationally intensive and require labels for every image. We therefore developed a new simple, cheap and scalable self-supervised pruning metric that demonstrates comparable performance to the best supervised metrics. Overall, our work suggests that the discovery of good data-pruning metrics may provide a viable path forward to substantially improved neural scaling laws, thereby reducing the resource costs of modern deep learning.

1 Introduction

Empirically observed neural scaling laws [1, 2, 3, 4, 5, 6, 7] in many domains of machine learning, including vision, language, and speech, demonstrate that test error often falls off as a power law with either the amount of training data, model size, or compute. Such power law scaling has motivated significant societal investments in data collection, compute, and associated energy consumption. However, power law scaling is extremely weak and unsustainable. For example a drop in error from 3% to 2% might require an *order of magnitude* more data, compute, or energy. For example, in language modeling with large transformers, a drop in cross entropy loss from about 3.4 nats to 2.8 requires *10 times* more training data (Fig. 1 in [2]). Also, for large vision transformers, an additional 2 *billion* pre-training data points (starting from 1 billion) leads to an accuracy gain on ImageNet of a few percentage points (Fig. 1 in [6]). Here we ask whether we might be able to do better. For example, can we achieve exponential scaling instead, with a good strategy for selecting training examples? Such vastly superior scaling would mean that we could go from 3% to 2% error by only adding a few carefully chosen training examples, rather than collecting $10\times$ more random ones.

Here we focus on scaling of performance with training dataset size, and demonstrate that exponential scaling is possible, both in theory and practice. The key idea is that power law scaling of error with respect to data suggests that many training examples are highly redundant. Thus one should in principle be able to prune training datasets to much smaller sizes and train on the smaller pruned

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

datasets without sacrificing performance. Indeed some recent works [8, 9, 10] have demonstrated this possibility by suggesting various metrics to sort training examples in order of their difficulty or importance, ranging from easy or redundant examples to hard or important ones, and pruning datasets by retaining some fraction of the hardest examples. However, these works leave open fundamental theoretical and empirical questions: When and why is successful data pruning possible? What are good metrics and strategies for data pruning? Can such strategies beat power law scaling? Can they scale to ImageNet? Can we leverage large *unlabeled* datasets to successfully prune labeled datasets? We address these questions through both theory and experiment. Our main contributions are:

1. Employing statistical mechanics, we develop a new analytic theory of data pruning in the student-teacher setting for perceptron learning, where examples are pruned based on their teacher margin, with large (small) margins corresponding to easy (hard) examples. Our theory quantitatively matches numerical experiments and reveals two striking predictions:
   (a) The optimal pruning strategy changes depending on the amount of initial data; with abundant (scarce) initial data, one should retain only hard (easy) examples.
   (b) Exponential scaling is possible with respect to pruned dataset size provided one chooses an increasing Pareto optimal pruning fraction as a function of initial dataset size.
2. We show that the two striking predictions derived from theory hold also in practice in much more general settings. Indeed we empirically demonstrate signatures of exponential scaling of error with respect to pruned dataset size for ResNets trained from scratch on SVHN, CIFAR-10 and ImageNet, and Vision Transformers fine-tuned on CIFAR-10.
3. Motivated by the importance of finding good quality metrics for data pruning, we perform a large scale benchmarking study of 10 different data pruning metrics at scale on ImageNet, finding that most perform poorly, with the exception of the most compute intensive metrics.
4. We leveraged self-supervised learning (SSL) to developed a new, cheap *unsupervised* data pruning metric that does *not* require labels, unlike prior metrics. We show this unsupervised metric performs comparably to the best supervised pruning metrics that require labels and much more compute. This result opens the door to the exciting possibility of leveraging pre-trained foundation models to prune new datasets even before they are labeled.

Overall these results shed theoretical and empirical insights into the nature of data in deep learning and our ability to prune it, and suggest our current practice of collecting extremely large datasets may be highly inefficient, despite its power law scaling. Our initial results in beating power-law scaling motivate further studies and investments in not just inefficently collecting large amounts of random data, but rather, intelligently collecting much smaller amounts of carefully selected data, leading to the creation and dissemination of *foundation datasets*, in addition to foundation models [11].

2 Background and related work

Our work brings together 3 largely disparate strands of intellectual inquiry in machine learning: (1) explorations of different metrics for quantifying differences between individual training examples; (2) the empirical observation of neural scaling laws; and (3) the statistical mechanics of learning.

2.1 Pruning metrics: not all training examples are created equal

Several recent works have explored various metrics for quantifying individual differences between data points. To describe these metrics in a uniform manner, we will think of all of them as ordering data points by their difficulty, ranging from "easiest" to "hardest." When these metrics have been used for data pruning, the hardest examples are retained, while the easiest ones are pruned away.

EL2N scores. For example [9] trained very small ensembles (of about 10) networks for a very short time (about 10 epochs) and computed for every training example the average L_2 norm of the error vector (EL2N score). Data pruning by retaining only the hardest examples with largest error enabled training from scratch on only 50% and 75% of CIFAR-10 and CIFAR-100 respectively without any loss in final test accuracy. However the performance of EL2N on ImageNet has not yet been explored.

Forgetting scores and classification margins. [8] noticed that over the entire course of training, some examples are learned early and never forgotten, while others can be learned and unlearned (i.e.

forgotten) repeatedly. They developed a forgetting score which measures the degree of forgetting of each example. Intuitively examples with low (high) forgetting scores can be thought of as easy (hard) examples. [8] explored data pruning using these metrics, but not at ImageNet scale.

Memorization and influence. [12] defined a memorization score for each example, corresponding to how much the probability of the correct label for the example increases when it is present in the training set relative to when it is absent; a large increase means the example must be memorized (i.e. the remaining training data do not suffice to correctly learn this example). Additionally [12] also considered an influence score that quantifies how much adding a particular example to the training set increases the probability of the correct class label of a test example. Intuitively, low memorization and influence scores correspond to easy examples that are redundant with the rest of the data, while high scores correspond to hard examples that must be individually learned. [12] did not use these scores for data pruning as their computation is expensive. We note since memorization explicitly approximates the increase in test loss due to removing each individual example, it is likely to be a good pruning metric (though it does not consider interactions when removing more than one).

Ensemble active learning. Active learning iterates between training a model and selecting new inputs to be labeled [13, 14, 15, 16, 17]. In contrast, we focus on data pruning: one shot selection of a data subset sufficient to train to high accuracy from scratch. A variety of coreset algorithms (e.g. [18]) have been proposed for this, but their computation is expensive, and so data-pruning has been less explored at scale on ImageNet. An early clustering approach [19] allowed training on 90% of ImageNet without sacrificing accuracy. Notably [10] reduced this to 80% by training a large ensemble of networks on ImageNet and using ensemble uncertainty to define the difficulty of each example, with low (high) uncertainty corresponding to easy (hard) examples. We will show how to achieve similar pruning performance without labels or the need to train a large ensemble.

Diverse ensembles (DDD). [20] assigned a score to every ImageNet image, given by the number of models in a diverse ensemble (10 models) that misclassified the image. Intuitively, low (high) scores correspond to easy (hard) examples. The pruning performance of this metric remains unexplored.

Summary. We note: (1) only one of these metrics has tested well for its efficacy in data pruning at scale on ImageNet; (2) *all* of these metrics require label information; (3) there is no theory of when and why data pruning is possible for any of these metrics; and (4) none of these works suggest the possibility of exponential scaling. We thus go beyond this prior work by benchmarking the data pruning efficacy of not only these metrics but also new unsupervised metrics we introduce that do not require label information, all at scale on ImageNet. We also develop an analytic theory for data-pruning for the margin metric that predicts not only the possibility of exponential scaling but also the novel finding that retaining easy, not hard examples is better when data is scarce.

2.2 Neural scaling laws and their potential inefficiency

Recent work [1, 2, 3, 4, 5, 7] has demonstrated that test loss $\mathcal{L}$ often falls off as a power law with different resources like model parameters (N), number of training examples (P), and amount of compute (C). However, the exponents ν of these power laws are often close to 0, suggesting potentially inefficient use of resources. For example, for large models with lots of compute, so that the amount of training data constitutes a performance bottleneck, the loss scales as $\mathcal{L} \approx P^{-\nu}$. Specifically for a large transformer based language model, $\nu = 0.095$, which implies *an order of magnitude* increase in training data drops cross-entropy loss by only about 0.6 nats (Fig. 1 in [2]). In neural machine translation experiments ν varies across language pairs from 0.35 to 0.48 (Table 1 in [4]). Interestingly, [7] explored a fixed computation budget C and optimized jointly over model size N and training set size P, revealing that scaling both N and P commensurately as C increases is compute optimal, and can yield smaller high performing models (trained on more data) than previous work. Nevertheless, for a transformer based language model, a $100\times$ increase in compute, corresponding to $10\times$ increases in *both* model size and training set size, leads to a drop in cross-entropy loss of about 0.5 nats (Fig 2 in [7]). Similar slow scaling holds for large vision transformers where adding 2 billion pre-training images reduces ImageNet performance by a few percentage points (Fig. 1 in [6]). While all of these results constitute significant improvements in performance, they do come at a substantial resource cost whose fundamental origin arises from power law scaling with small exponents. Here, in the specific case of data scaling, we explore whether we can beat power law scaling through careful data selection.

2.3 Statistical mechanics of perceptron learning

Statistical mechanics has long played a role in analyzing machine learning problems (see e.g. [21, 22, 23, 24] for reviews). One of the most fundamental applications is perceptron learning in the student-teacher setting [25, 26], in which random i.i.d. Gaussian inputs are labeled by a teacher perceptron to construct a training set. The test error for another student perceptron learning from this training set then scales as a power law with exponent -1 for such data. Such perceptrons have also been analyzed in an active learning setting where the learner is free to design *any* new input to be labeled [27, 28], rather than choose from a fixed set of inputs, as in data-pruning. Recent work [29] has analyzed this scenario but focused on message passing algorithms that are tailored to the case of Gaussian inputs and perceptrons, and are hard to generalize to real world settings. In contrast we analyze margin based pruning algorithms that are used in practice in diverse settings, as in [8, 9].

3 An analytic theory of data pruning

To better understand data pruning, we employed the replica method from statistical mechanics [30] to develop an analytic theory of pruning for the perceptron in the student-teacher setting [21] (see App. for detailed derivations of all results). Consider a training dataset of P examples $\{\mathbf{x}^\mu, y^\mu\}_{\mu=1,\ldots,P}$ where $\mathbf{x}^\mu \in \mathbb{R}^N$ are i.i.d. zero mean unit variance random Gaussian inputs and $y^\mu = \text{sign}(\mathbf{T} \cdot \mathbf{x}^\mu)$ are labels generated by a teacher perceptron with weight vector $\mathbf{T} \in \mathbb{R}^N$. We work in the high dimensional statistics limit where $N, P \to \infty$ but the ratio $\alpha_{\text{tot}} = \frac{P}{N}$ of the number of total training examples to parameters remains $O(1)$. We then consider a pruning algorithm used in [8, 9], namely: (1) train a probe student perceptron for very few epochs on the training data, obtaining weights $\mathbf{J}_{\text{probe}}$; (2) compute the margin $m^\mu = \mathbf{J}_{\text{probe}} \cdot (y^\mu \mathbf{x}^\mu)$ of each training example, where large (small) margins correspond to easy (hard) examples; (3) construct a pruned dataset of size $P_{\text{prune}} = fP$, where f is the fraction of examples kept, by retaining the P_{prune} hardest examples, (4) train a new perceptron to completion on the smaller dataset with a smaller ratio $\alpha_{\text{prune}} = \frac{P_{\text{prune}}}{N}$ of examples to parameters.

We are interested in the test error ε of this final perceptron as a function of α_{tot}, f, and the angle θ between the probe student $\mathbf{J}_{\text{probe}}$ and the teacher $\mathbf{T}$. Our theory approximates $\mathbf{J}_{\text{probe}}$ as simply a random Gaussian vector conditioned to have angle θ with the teacher $\mathbf{T}$. Under this approximation we obtain an analytic theory for $\varepsilon(\alpha_{\text{tot}}, f, \theta)$ that is asymptotically exact in the high dimensional limit (App.). We first examine results when $\theta = 0$, so we are pruning training examples according to their veridical margins with respect to the teacher (Fig 1A). We find two striking phenomena, each of which constitute predictions in real-world settings that we will successfully confirm.

The best pruning strategy depends on the amount of initial data. First, we note the test error curve for $f = 1$ in Fig 1A corresponding to no pruning, or equivalently to *randomly* pruning a larger dataset of size α_{tot} down to a size α_{prune}, exhibits the well known classical perceptron learning power law scaling $\varepsilon \propto \alpha_{\text{prune}}^{-1}$. Interestingly though, for small α_{tot}, keeping the hardest examples performs *worse* than random pruning (lighter curves above darkest curve for small α_{prune} in Fig 1A). However, for large α_{tot}, keeping the hardest examples performs *substantially better* than random pruning (lighter curves below darkest curve for small α_{prune} in Fig 1A). It turns out keeping the *easiest* rather than hardest examples is a better pruning strategy when α_{tot} is small (Fig 1C). Intuitively, if one does not have much data to start with, it is better to keep the easiest examples with largest margins (i.e. the blue regions of Fig 1B) to avoid overfitting. Otherwise, if one starts with lots of data, so overfitting is less of an issue, it is best to keep the hardest examples with smallest margin that provide more information about the teacher's decision boundary (i.e. the green region of Fig 1B). Fig 1C reveals which pruning strategy is best as a joint function of α_{tot} and f. Note the transition between optimal strategies becomes sharper at small fractions f of data kept. This transition between optimal pruning strategies can be viewed as a prediction in more general settings. To test this prediction we trained a ResNet18 on pruned subsets of the CIFAR-10 dataset (Fig. 1D), and observed strikingly similar behavior, indicating the prediction can hold far more generally, beyond perceptron learning. Interestingly, [8, 9] missed this transition, likely because they started pruning from large datasets.

Pareto optimal data pruning can beat power law scaling. A second prediction of our theory is that when keeping a *fixed* fraction f of the hardest examples as α_{tot} increases (i.e. constant color curves in Fig 1A), the error initially drops exponentially in $\alpha_{\text{prune}} = f\alpha_{\text{tot}}$, but then settles into the universal power law $\varepsilon \propto \alpha_{\text{prune}}^{-1}$ for all fixed f. Thus there is no asymptotic advantage to data pruning

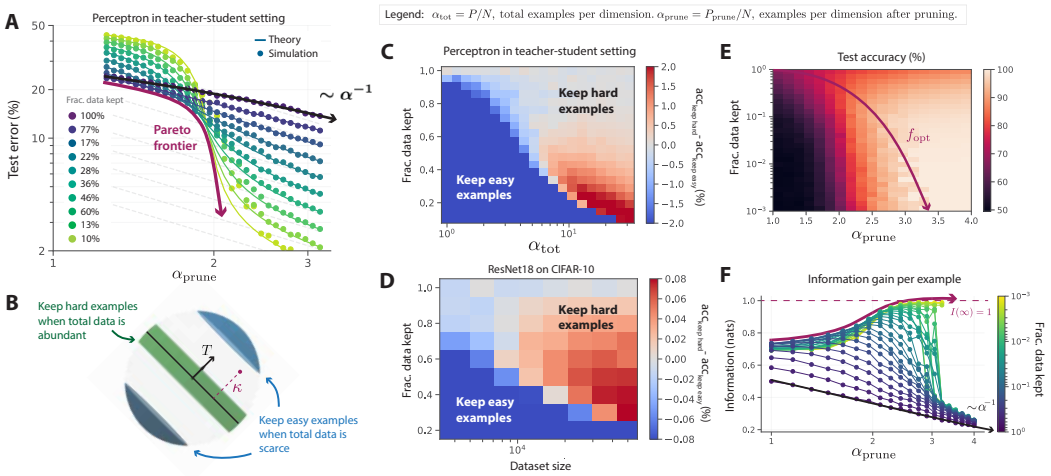

Figure 1: Our analytic theory of data pruning predicts that power law scaling of test error with respect to dataset size can be beaten. **A:** Test error as a function of $\alpha_{\text{prune}} = f\alpha_{\text{tot}}$ with $\theta = 0$. We observe an excellent match between our analytic theory (solid curves) and numerical simulations of perceptron learning at N=200 (dots). The red curve indicates the Pareto optimal test error ε achievable from a tradeoff between α_{tot} and f at fixed α_{prune}. **B:** We find that when data is abundant (scarce) corresponding to large (small) α_{tot}, the better pruning strategy is to keep the hard (easy) examples. **C:** Color indicates difference in test error in keeping hard versus easy examples, revealing the change in strategy in (B). **D:** We tested this prediction on a ResNet18 trained on CIFAR10, finding remarkably the same shift in optimal pruning strategy under the EL2N metric. **E:** Test accuracy as a function of f and α_{prune}. For every fixed α_{prune}, there is an optimal f_{opt} (blue curve). **F:** $I(\alpha_{\text{prune}})$ for different f.

at a fixed f. However, by pruning more aggressively (smaller f) when given more initial data (larger α_{tot}), one can achieve a Pareto optimal test error as a function of pruned dataset size α_{prune} that remarkably traces out at least an exponential scaling law (Fig. 1C, red curve). Indeed our theory predicts for each α_{prune} a Pareto optimal point in α_{tot} and f (subject to $\alpha_{\text{prune}} = f\alpha_{\text{tot}}$), yielding for every fixed α_{prune} an optimal f_{opt}, plotted in Fig. 1E. Note f_{opt} decreases with α_{prune} indicating more aggressive pruning (smaller f_{opt}) of original datasets of larger size α_{tot} is required to obtain larger Pareto optimal pruned datasets of size α_{prune}. We will test this striking scaling prediction in Fig 3.

Beating power law scaling: an information-theoretic perspective. Classical randomly selected data generates slow power law error scaling because each extra training example provides less new information about the correct decision boundary than the previous example. More formally, let $S(\alpha_{\text{tot}})$ denote the typical entropy of the posterior distribution over student perceptron weights consistent with a training set of size α_{tot}. The information gain $I(\alpha_{\text{tot}})$ due to additional examples beyond α_{tot} can be defined as the rate at which the posterior entropy is reduced: $I(\alpha_{\text{tot}}) = -\frac{d}{d\alpha_{\text{tot}}} S(\alpha_{\text{tot}})$. In classical perceptron learning $I(\alpha_{\text{tot}})$ decays to zero as a power law in α_{tot}, reflecting a vanishing amount of information per each new example, leading to the slow power law decay of test error $\varepsilon \propto \alpha_{\text{tot}}^{-1}$.

However, data pruning can increase the information gained per example by pruning away the uninformative examples. To show this, we generalized the replica calculation of the posterior entropy S and information gain I from random datasets of size α_{tot} to pruned datasets of size α_{prune} (App.). We plot the resulting information gain $I(\alpha_{\text{prune}})$ for different f in Fig. 1F. For any fixed f, $I(\alpha_{\text{prune}})$ will eventually decay as a power law as $\alpha_{\text{prune}}^{-1}$. However, by more aggressively pruning (smaller f) datasets of larger size α_{tot}, $I(\alpha_{\text{prune}})$ can converge to a finite value $I(\infty) = 1$ nat/example, resulting in larger pruned datasets only adding useful non-redundant information. Since each new example under Pareto optimal data pruning conveys finite information about the target decision boundary, as seen in Fig. 1F, the test error can decay at least exponentially in pruned dataset size as in Fig. 1A.

An imperfect pruning metric yields a cross over from exponential to power law scaling. We next examine the case of nonzero angle θ between the probe student $\mathbf{J}_{\text{probe}}$ and the teacher $\mathbf{T}$, such

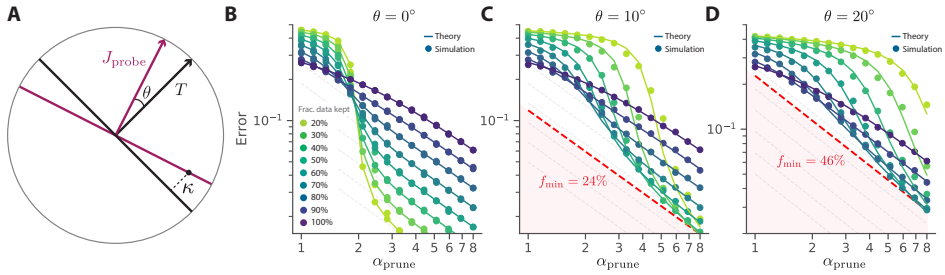

Figure 2: Data pruning with an imperfect metric. **A:** Weight vectors and decision boundaries for a teacher (black) and probe student (red) separated by angle θ. The black point has margin 0 (κ) w.r.t. the probe (teacher). **B–D:** Test error as a function of α_{prune} for different f and different θ.

that the ranking of training examples by margin is no longer completely accurate (Fig. 2A). Retaining the hard examples with smallest margin with respect to the probe student will always result in pruned datasets lying near the probe's decision boundary. But if θ is large, such examples might be far from the teacher's decision boundary, and therefore could be less informative about the teacher (Fig. 2A). As a result our theory, confirmed by simulations, predicts that under nonzero angles θ, the Pareto optimal lower envelope of test error over both α_{tot} and f initially scales exponentially as a function of $\alpha_{\text{prune}} = f\alpha_{\text{tot}}$ but then crosses over to a power-law (Fig 2BCD). Indeed, at any given nonzero θ, our theory reveals that as α_{tot} (and therefore α_{prune}) becomes large, one cannot decrease test error any further by retaining less than a minimum fraction $f_{\min}(\theta)$ of all available data. For example when $\theta = 10°$ ($\theta = 20°$) one can do no better asymptotically than pruning down to 24% (46%) of the total data (Fig. 2CD). As θ approaches 0, $f_{\min}(\theta)$ approaches 0, indicating that one can prune extremely aggressively to arbitrarily small f while still improving performance, leading to at least exponential scaling for arbitrarily large α_{prune} in Fig. 2B. However, for nonzero θ, the lack of improvement for $f < f_{\min}(\theta)$ at large α_{prune} renders aggressive pruning ineffective. This result highlights the importance of finding high quality pruning metrics with $\theta \approx 0$. Such metrics can delay the cross over from exponential to power law scaling as pruned dataset size α_{prune} increases, by making aggressive pruning with very small f highly effective. Strikingly, in App. Fig 7 we demonstrate this cross-over in a real-world setting by showing that the test error on SVHN is bounded below by a power law when the dataset is pruned by a probe ResNet18 under the EL2N metric, trained for 4 epochs (weak pruning metric) but not a probe ResNet18 trained for 40 epochs (strong pruning metric).

4 Data pruning can beat power law scaling in practice

Our theory of data pruning for the perceptron makes three striking predictions which can be tested in more general settings, such as deep neural networks trained on benchmark datasets: (1) relative to random data pruning, keeping only the hardest examples should *help* when the initial dataset size is large, but *hurt* when it is small; (2) data pruning by retaining a fixed fraction f of the hardest examples should yield power law scaling, with exponent equal to that of random pruning, as the initial dataset size increases; (3) the test error optimized over both initial data set size and fraction of data kept, can trace out a Pareto optimal lower envelope that beats power law scaling of test error as a function of pruned dataset size, through more aggressive pruning at larger initial dataset size. We verified all three of these predictions on ResNets trained on SVHN, CIFAR10, and ImageNet using varying amounts of initial dataset size and fractions of data kept under data pruning (compare theory in Fig 3A with deep learning experiments in Fig 3BCD). In each experimental setting we see better than power-law scaling at larger initial data set sizes and more aggressive pruning. Moreover we would likely see even better scaling with even larger initial datasets (as in Fig 3A dashed lines).

Data pruning for transfer. Modern foundation models are pretrained on a large initial dataset, and then transferred to other downstream tasks by fine-tuning on them. We therefore also examined whether data-pruning can be effective for both reducing the amount of fine-tuning data and the amount of pre-training data. To this end, we first analyzed a vision transformer (ViT) pre-trained on ImageNet21K and then fine-tuned on different pruned subsets of CIFAR-10. Interestingly, pretrained models allow for far more aggressive data pruning; fine-tuning on only 10% of CIFAR-10 can match or exceed performance obtained by fine tuning on *all* of CIFAR-10 (Fig. 4A). Furthermore Fig. 4A

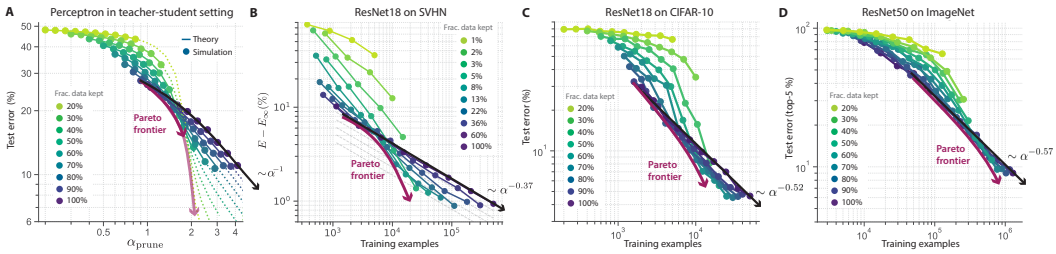

Figure 3: Beating power law scaling in practice. **A–D:** Curves of test error against pruned dataset size in 4 settings. Pruning scores were EL2N [9] for CIFAR-10 and SVHN and memorization [12] for ImageNet. See App. B for all pruning/training details and App C for similar ImageNet plots with EL2Ns. Note solid curves reflect performance with a fixed total dataset size; if we prune more aggressively with even larger datasets, scaling could improve further (e.g., dashed lines in **A**). Error curves with no data pruning ($f = 1$) are labeled with their best-fit power-law scaling $\sim \alpha^{-\nu}$. For SVHN in **B** we subtract an asymptotic constant error $E(P \to \infty) = 1.1\%$ from each of the curves to reveal more clearly the power law scaling.

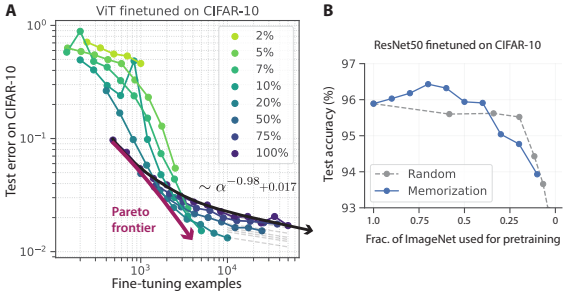

Figure 4: Data pruning for transfer. **A**: CIFAR10 performance of a ViT pre-trained on all of ImageNet21K and fine-tuned on different pruned subsets of CIFAR10 under the EL2N metric. **B:** CIFAR10 performance of ResNet50s pre-trained on different pruned subsets of ImageNet1K and fine-tuned on all of CIFAR10.

provides a new example of beating power-law scaling in the setting of fine-tuning. Additionally, we examined the efficacy of pruning pre-training data by pre-training ResNet50s on different pruned subsets of ImageNet1K (exactly as in Fig. 3D) and then fine-tuning them on all of CIFAR10. Fig. 4B demonstrates pre-training on as little as 50% of ImageNet can match or exceed CIFAR10 performance obtained by pre-training on all of ImageNet. Thus intriguingly pruning pre-training data on an upstream task can still maintain high performance on a different downstream task. Overall these results demonstrate the promise of data pruning in transfer-learning for both the pre-training and fine-tuning phases.

5 Benchmarking supervised pruning metrics on ImageNet

We note that the majority of data pruning experiments have been performed on small-scale datasets (i.e. variants of MNIST and CIFAR), while the few pruning metrics proposed for ImageNet have rarely been compared against baselines designed on smaller datasets. Therefore, it is currently unclear how most pruning methods scale to ImageNet and which method is best. Motivated by how strongly the quality of a pruning metric can impact performance in theory (Fig. 2), we decided to fill this knowledge gap by performing a systematic evaluation of 8 different supervised pruning metrics on ImageNet: two variants of influence scores [12], two variants of EL2N [9], DDD [20], memorization [12], ensemble active learning [10], and forgetting [8]. See Section 2 for a review of these metrics. Additionally, we include two new prototypicality metrics that we introduce in the next section.

We first asked how consistent the rankings induced by different metrics are by computing the Spearman rank correlation between each pair of metrics (Figure 5A). Interestingly, we found substantial diversity across metrics, though some (EL2N, DDD, and memorization) were fairly similar with rank correlations above 0.7. However, we observed marked performance differences between metrics: Fig 5BC shows test performance when a fraction f of the hardest examples under each metric are kept in the training set. Despite the success of many of these metrics on smaller datasets, only a few still match performance obtained by training on the full dataset, when selecting a significantly smaller training subset (i.e. about 80% of ImageNet). Nonetheless, most metrics continue to beat random pruning, with memorization in particular demonstrating strong performance (Fig 5C).

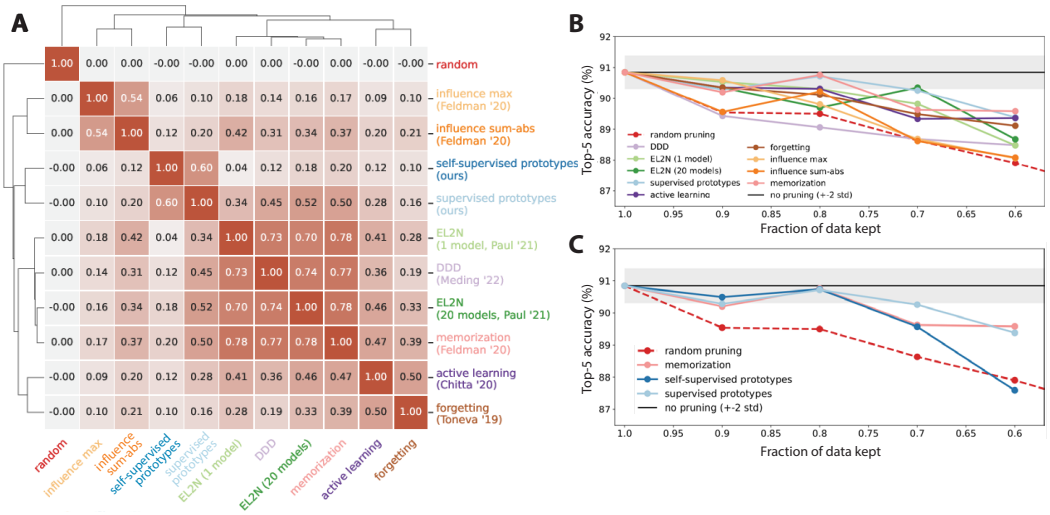

Figure 5: Dataset pruning at ImageNet scale. **A:** Spearman's rank correlation between all pairs of ImageNet metric scores, along with hierarchical clustering (as provided by `seaborn.clustermap`). **B:** Benchmarking supervised metrics on ImageNet (top-5 validation accuracy). **C:** Comparing top-5 performance on ImageNet when pruning according to the best existing supervised metric (memorization) and our supervised and self-supervised prototype metrics. In all 3 cases, training on 80% of ImageNet matches training on 100%. See App. for pruning and training details.

We found that all pruning metrics amplify class imbalance, which results in degraded performance. To solve this we used a simple 50% class balancing ratio for all ImageNet experiments. Further details and baselines without class balancing are shown in App. G.

6 Self-supervised data pruning through a prototypicality metric

Figure 5 shows many data pruning metrics do not scale well to ImageNet, while the few that do require substantial amounts of compute. Furthermore, all these metrics require labels, thereby limiting their ability to prune data for large-scale foundation models trained on massive unlabeled datasets [11]. Thus there is a clear need for simple, scalable, self-supervised pruning metrics.

To compute a self-supervised pruning metric for ImageNet, we perform k-means clustering in the embedding space of an ImageNet pre-trained self-supervised model, such as SWaV [31], and define the difficulty of each data point by the distance to its nearest cluster centroid, or prototype. Thus easy (hard) examples are the most (least) prototypical. Encouragingly, in Figure 5C, we find our self-supervised prototype metric matches or exceeds the performance of the best supervised metric, memorization, until only 70% of the data is kept, despite the fact that our metric does not use labels and is much simpler and cheaper to compute than many previously proposed supervised metrics. See App. Fig. 8 for further scaling experiments using the self-supervised metric.

To assess whether the clusters found by our metric align with ImageNet classes, we compared their overlaps in Figure 6A. Interestingly, we found alignment for some but not all classes. For example, class categories such as snakes were largely aligned to a small number of unsupervised clusters, while other classes were dispersed across many such clusters. If class information is available, we can enforce alignment between clusters and classes by simply computing a single prototype for each class (by averaging the embeddings of all examples of this class). While originally intended to be an additional baseline metric (called supervised prototypes, light blue in Fig 5BC), this metric remarkably outperforms all other supervised metrics and largely matches the performance of memorization. Moreover, when more than 70% of the data are kept, the performance of the self-supervised and supervised metrics are similar, demonstrating the promise of self-supervised pruning.

One important choice for the self-supervised prototype metric is the number of clusters k. We found, reassuringly, our results were robust to this choice: k can deviate one order of magnitude more or less than the true number of classes (i.e. 1000 for ImageNet) without affecting performance (App. E).

To better understand example difficulty under various metrics, we visualize extremal images for our self-supervised prototype metric and the memorization metric for one class (Fig 6B,C). Qualitatively, easy examples correspond to highly similar, redundant images, while hard examples look like idiosyncratic outliers. See App. D, Figures 9,10,11,12,13,14,15,16 for more classes and metrics.

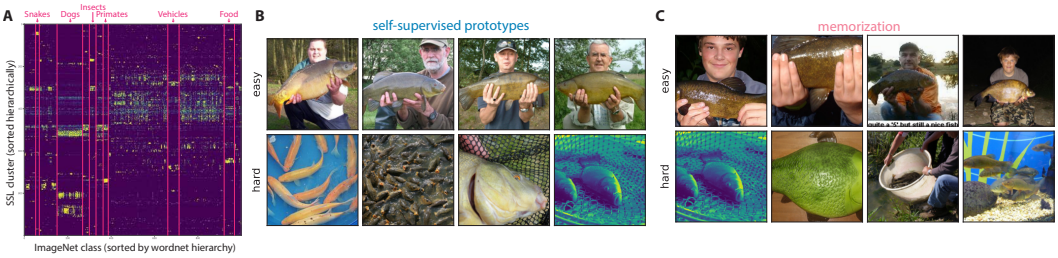

Figure 6: **A:** Heat map where each row denotes the probability images in a given cluster come from each ImageNet class. **B:** Examples of the easiest and hardest images under our self-supervised pruning metric and the best supervised metric (memorization, shown in **C**) for ImageNet class 0 (`tench`). Note images in B and C exhibit some of the biases inherent in ImageNet [e.g. 32, 33].

7 Discussion

Summary. We have shown, both in theory and practice, how to break beyond slow power-law scaling of error versus dataset size to faster exponential scaling, through data pruning. Additionally we have developed a simple self-supervised pruning metric that enables us to discard 20% of ImageNet without sacrificing performance, on par with the best and most compute intensive supervised metric.

Limitations. The most notable limitation is that achieving exponential scaling requires a high quality data pruning metric. Since most metrics developed for smaller datasets scale poorly to ImageNet, our results emphasize the importance of future work in identifying high quality, scalable metrics. Our self-supervised metric provides a strong initial baseline.

Also a key advantage of data pruning is reduced computational cost due to training on a smaller dataset for the same number of epochs as the full dataset (provided the computational cost of computing the pruning metric does not dominate). However, we found that performance often increased when the same number training *iterations* on the full dataset was carried out on the pruned dataset, resulting in the same training time, but additional training epochs. However, this performance gain saturated *before* training time on the pruned dataset approached that on the whole dataset (App. I) still yielding computational efficiency gain. Overall this tradeoff between accuracy and training time on pruned data is important to consider in evaluating potential gains due to data pruning.

Finally, we found that class-balancing (whether via supervision or unsupervised clusters) was essential to maintain performance on data subsets (App. G). Future work will be required to identify ways to effectively select the appropriate amount of class-balancing.

Ethical considerations. A potential negative societal impact could be that data-pruning leads to unfair outcomes for certain groups. We have done a preliminary analysis of how data-pruning affects performance on individual ImageNet classes H, finding no substantial differential effects across classes. However proper fairness tests specific to deployment settings should always be conducted on every model, whether trained on pruned data or not.

Outlook: Towards foundation datasets. We believe the most promising future direction is the further development of scalable, unsupervised data pruning metrics. Indeed our theory predicts that the application of pruning metrics on larger scale datasets should yield larger gains by allowing more aggressive pruning. This makes data pruning especially exciting for use on the massive unlabeled datasets used to train large foundation models (e.g. 400M image-text pairs for CLIP [34], 3.5B Instagram images [35], 650M images for the DALLE-2 encoder [36], 780B tokens for PALM [37]). If highly pruned versions of these datasets can be used to train a large number of different models, one can even conceive of such carefully chosen data subsets as *foundation datasets* in which the initial computational cost of data pruning can be amortized across efficiency gains in training many downstream models, just at the initial computational cost of training foundation models is amortized across the efficiency gains of fine-tuning across many downstream tasks. Together, our results demonstrate the promise and potential of data pruning for large-scale training and pretraining.

References

[1] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. December 2017.

[2] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. January 2020.

[3] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling. October 2020.

[4] Mitchell A Gordon, Kevin Duh, and Jared Kaplan. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5915–5922, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[5] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. February 2021.

[6] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. June 2021.

[7] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal large language models. March 2022.

[8] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. In *ICLR*, 2019.

[9] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Adv. Neural Inf. Process. Syst.*, 34, December 2021.

[10] Kashyap Chitta, José M Álvarez, Elmar Haussmann, and Clément Farabet. Training data subset search with ensemble active learning. *IEEE Trans. Intell. Transp. Syst.*, pages 1–12, 2021.

[11] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W Thomas, Florian Tramèr, Rose E Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. August 2021.

[12] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Adv. Neural Inf. Process. Syst.*, 33:2881–2891, 2020.

[13] Burr Settles. Active learning literature survey. 2009.

[14] Antoine Bordes, Seyda Ertekin, Jason Weston, Léon Botton, and Nello Cristianini. Fast kernel classifiers with online and active learning. *J. Mach. Learn. Res.*, 6(9), 2005.

[15] Zeyad Ali Sami Emam, Hong-Min Chu, Ping-Yeh Chiang, Wojciech Czaja, Richard Leapman, Micah Goldblum, and Tom Goldstein. Active learning at the ImageNet scale. November 2021.

[16] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A Core-Set approach. August 2017.

[17] Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher D Manning. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. July 2021.

[18] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In Hal Daumé Iii and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6950–6960. PMLR, 2020.

[19] V Birodkar, H Mobahi, and S Bengio. Semantic redundancies in Image-Classification datasets: The 10% you don't need. *arXiv preprint arXiv:1901.11409*, 2019.

[20] Kristof Meding, Luca M. Schulze Buschoff, Robert Geirhos, and Felix A. Wichmann. Trivial or impossible—dichotomous data difficulty masks model differences (on ImageNet and beyond). In *International Conference on Learning Representations*, 2022.

[21] A Engel and C V den Broeck. *Statistical Mechanics of Learning*. Cambridge Univ. Press, 2001.

[22] Madhu Advani, Subhaneil Lahiri, and Surya Ganguli. Statistical mechanics of complex neural systems and high dimensional data. *J. Stat. Mech: Theory Exp.*, 2013(03):P03014, 2013.

[23] Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, March 2020.

[24] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: thresholds and algorithms. *Adv. Phys.*, 65(5):453–552, September 2016.

[25] E Gardner. The space of interactions in neural network models. *J. of Physics A*, 21:257–270, 1988.

[26] H S Seung, H Sompolinsky, and N Tishby. Statistical mechanics of learning from examples. *Phys. Rev. A*, 45(8):6056, 1992.

[27] Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Information, prediction, and query by committee. *Adv. Neural Inf. Process. Syst.*, 5, 1992.

[28] Hai-Jun Zhou. Active online learning in the binary perceptron problem. *Commun. Theor. Phys.*, 71(2):243, February 2019.

[29] H Cui, L Saglietti, and L Zdeborová. Large deviations in the perceptron model and consequences for active learning, 2021.

[30] M Mezard, G Parisi, and M A Virasoro. *Spin glass theory and beyond*. World scientific Singapore, 1987.

[31] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

[32] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558, 2020.

[33] Yuki M Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. Pass: An imagenet replacement for self-supervised pretraining without humans. *arXiv preprint arXiv:2109.13228*, 2021.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[35] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[37] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[38] Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. VISSL. https://github.com/facebookresearch/vissl, 2021.

[39] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section...
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See discussion section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See discussion section 7 and App. H.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See App. A
   (b) Did you include complete proofs of all theoretical results? [Yes] See App. A
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All code for theory plots and numerical simulations is packaged in a reproducible colab notebook to be released upon acceptance. All code for training and evaluating deep neural networks will be placed in a github repo upon acceptance.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See App. B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Since we train a very large (order 100) number of models on computationally intensive tasks (e.g. ImageNet), many of our plots contain only results from a single random seed. However, variability due to random seeds can be inferred by the smooth progression of the relevant quantity on each plot.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See App. B
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See App. B
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See App. B
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]