TOWARDS ROBUST DATA PRUNING

Artem Vysogorets
Center for Data Science
New York University
amv458@nyu.edu

Julia Kempe
New York University
Meta FAIR

ABSTRACT

In the era of exceptionally data-hungry models, careful selection of the training data is essential to mitigate the extensive costs of deep learning. Data pruning offers a solution by removing redundant or uninformative samples from the dataset, which yields faster convergence and improved neural scaling laws. However, little is known about its impact on classification bias of the trained models. We conduct the first systematic study of this effect and reveal that existing data pruning algorithms can produce highly biased classifiers. At the same time, we argue that random data pruning with appropriate class ratios has potential to improve the worst-class performance. We propose a "fairness-aware" approach to pruning and empirically demonstrate its performance on standard computer vision benchmarks. In sharp contrast to existing algorithms, our proposed method continues improving robustness at a tolerable drop of average performance as we prune more from the datasets.

The ever-increasing state-of-the-art performance of deep learning models requires exponentially larger volumes of training data according to neural scaling laws (Hestness et al., 2017; Rosenfeld et al., 2020; Gordon et al., 2021). However, not all collected data is equally important for learning as it contains noisy, repetitive, or uninformative samples. A recent research thread on data pruning is concerned with removing those unnecessary data, resulting in improved convergence speed, neural scaling factors, and resource efficiency (Toneva et al., 2019; Paul et al., 2021; He et al., 2023). These methods design scoring mechanisms to assess the utility of each sample, often measured by its difficulty or uncertainty as approximated during a preliminary training round, that guides pruning. Sorscher et al. (2022) report that selecting high-quality data using these techniques can trace a Pareto optimal frontier, beating the notorious power scaling laws.

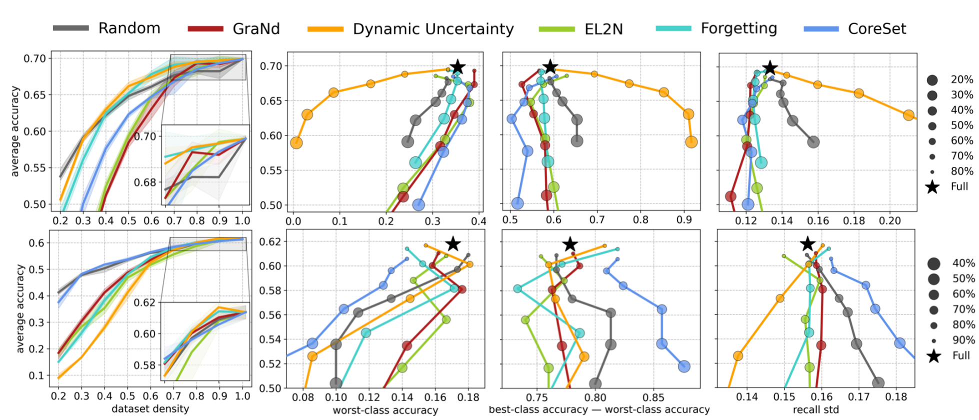

Figure 1: The average test performance of various data pruning algorithms against dataset density (fraction of samples remaining after pruning) and measures of class-wise fairness: worst-class accuracy, difference between best- and worst-class recall, and standard deviation of recall across classes. **Top:** VGG-19 on CIFAR-100, **Bottom:** ResNet-18 on TinyImageNet. All results averaged over 3 random seeds; error bands represent min/max.

The modus operandi of data pruning resembles algorithms that mitigate distributional bias—a well-established issue in AI systems concerning the performance disparity across protected groups of the population (e.g., race or gender) or classes (Dwork et al., 2012; Hardt et al., 2016). The early approaches in this domain improve the worst-class performance of imbalanced datasets by subsampling the majority classes (Barua et al., 2012; Cui et al., 2019; Tan et al., 2020). Data pruning extends this strategy by designing finer pruning criteria and working for balanced datasets, too. More recent developments apply importance weighting to emphasize under-represented or under-performing groups or classes during optimization (Sinha et al., 2022; Sagawa* et al., 2020; Liu et al., 2021; Idrissi et al., 2022; Wang et al., 2023; Lukasik et al., 2022; Chen et al., 2017). Data pruning adheres to this paradigm as well by effectively assigning zero weights to the removed training samples. These parallels indicate that data pruning has potential to reduce classification bias in deep learning models, all while offering greater resource and memory efficiency. We conduct the first systematic evaluation of various pruning algorithms with respect to distributional robustness and find that this potential remains largely unrealized (Figure 1). For example, Dynamic Uncertainty by He et al. (2023) achieves superior average test performance on CIFAR-100 with VGG-19 but fails miserably in terms of worst-class accuracy. Thus, it is imperative to benchmark pruning algorithms using a more comprehensive suite of performance metrics that reflect classification bias, and to develop solutions that address fairness directly.

In contrast to data pruning, many prior works on distributionally robust optimization compute difficulty-based importance scores at the level of groups or classes and not for individual training samples. For pruning, this would correspond to selecting appropriate class ratios but subsampling randomly within each class. To imitate this behavior, we propose to select class proportions based on the corresponding class-wise error rates computed on a hold-out validation set after a preliminary training on the full dataset, and call this procedure *Metric-based Quotas (MetriQ)*. In particular, given a target dataset density $d \in (0, 1)$, for each class $k \in [K]$ with recall r_k, we let its target density be $d_k = d(1 - r_k)/Z$ where Z is a normalizing factor. Note that this procedure may yield $d_k > 1$ for some $k \in [K]$ when d is sufficiently large; we do not prune such classes and redistribute the excess density across unsaturated ($d_k < 1$) classes according to their MetriQ proportions.

To validate the effectiveness of random pruning with MetriQ (Random+MetriQ), we compare it to various baselines derived from two of the strongest pruning algorithms: GraNd (Paul et al., 2021), and Forgetting (Toneva et al., 2019). In addition to plain random pruning, for each of these two strategies, we consider (1) random pruning that respects class-wise ratios automatically determined by the strategy (Random+StrategyQ), (2) applying the strategy for pruning within classes but distributing sparsity across classes according to MetriQ (Strategy+MetriQ), and (3) the strategy itself. As demonstrated in Figure 2, MetriQ improves robustness of the existing pruning algorithms and particularly shines when applied together with random pruning, substantially reducing the classification bias of models trained on the full dataset while offering an enhanced data efficiency.

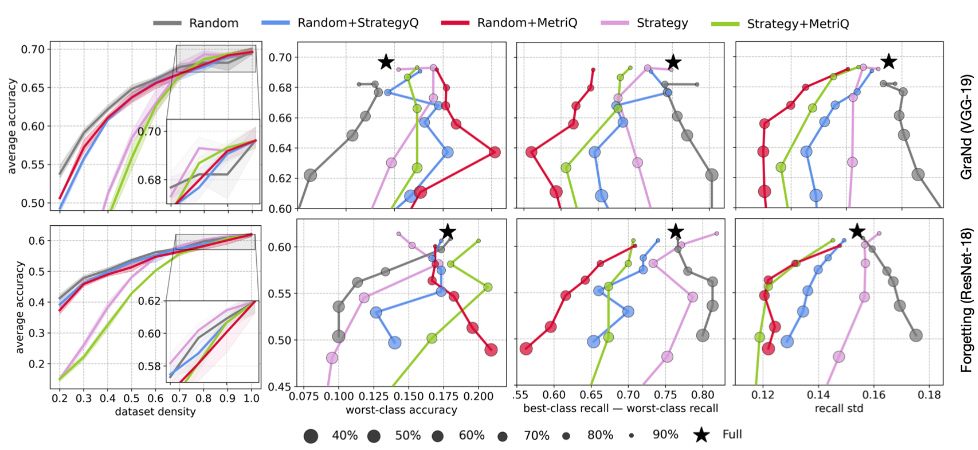

Figure 2: The average test performance of various data pruning protocols against dataset density and measures of class-wise fairness. **Top:** GraNd, VGG-19 on CIFAR-100; **Bottom:** Forgetting, ResNet-18 on TinyImageNet. All results averaged over 3 random seeds.

Acknowledgements. The first author was partly supported by the NSF Award 1922658. This work was supported in part through the NYU IT High Performance Computing resources, ser- vices, and staff expertise.

References

Sukarna Barua, Md Monirul Islam, Xin Yao, and Kazuyuki Murase. Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on knowledge and data engineering*, 26(2):405–425, 2012.

Robert S Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. Robust optimization for nonconvex objectives. *Advances in Neural Information Processing Systems*, 30, 2017.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. Data and parameter scaling laws for neural machine translation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5915–5922, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Muyang He, Shuo Yang, Tiejun Huang, and Bo Zhao. Large-scale dataset pruning with dynamic uncertainty. *arXiv preprint arXiv:2306.05175*, 2023.

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. *arXiv preprint arXiv:1712.00409*, 2017.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang (eds.), *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 336–351. PMLR, 11–13 Apr 2022.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 18–24 Jul 2021.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Teacher's pet: understanding and mitigating biases in distillation. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations (ICLR) 2020*, 2020.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Saptarshi Sinha, Hiroki Ohashi, and Katsuyuki Nakamura. Class-difficulty based methods for long-tailed visual recognition. *International Journal of Computer Vision*, 130(10):2517–2531, 2022.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11662–11671, 2020.

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 *International Conference on Learning Representations*, 2019.

Serena Wang, Harikrishna Narasimhan, Yichen Zhou, Sara Hooker, Michal Lukasik, and Aditya Krishna Menon. Robust distillation for worst-class performance: on the interplay between teacher and student objectives. In Robin J. Evans and Ilya Shpitser (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2237–2247. PMLR, 31 Jul–04 Aug 2023.