Does 'Deep Learning on a Data Diet' reproduce? Overall yes, but GraNd at Initialization does not

Anonymous authors
Paper under double-blind review

Abstract

The paper 'Deep Learning on a Data Diet' by Paul et al. (2021) introduces two innovative metrics for pruning datasets during the training of neural networks. While we are able to replicate the results for the EL2N score at epoch 20, the same cannot be said for the GraNd score at initialization. The GraNd scores later in training provide useful pruning signals, however. The GraNd score at initialization calculates the average gradient norm of an input sample across multiple randomly initialized models before any training has taken place. Our analysis reveals a strong correlation between the GraNd score at initialization and the input norm of a sample, suggesting that the latter could have been a cheap new baseline for data pruning. Unfortunately, neither the GraNd score at initialization nor the input norm surpasses random pruning in performance. This contradicts one of the findings in Paul et al. (2021). We were unable to reproduce their CIFAR-10 results using both an updated version of the original JAX repository and in a newly implemented PyTorch codebase. An investigation of the underlying JAX/FLAX code from 2021 surfaced a bug in the checkpoint restoring code that was fixed in April 2021[1].

1 Introduction

The rapidly growing field of deep learning has led to significant advances in various applications, including computer vision, natural language processing, and speech recognition. However, the immense amounts of data required to train these neural networks present challenges in terms of computational resources, storage, and energy consumption. As a result, there is a growing interest in finding methods to reduce data requirements while still maintaining model performance.

The senior author of 'Deep Learning on a Data Diet' (Paul et al., 2021) recently gave a talk at our lab that explored this issue, presenting their novel metrics for pruning datasets. During the talk, the author of this current work suggested *a correlation between the proposed GraNd score at initialization and input norms*, sparking further research into the effectiveness of these new pruning techniques. In this paper, we delve deeper into this intriguing question, exploring the practicality and efficacy of these metrics for data pruning.

'Deep Learning on a Data Diet'. Paul et al. (2021) introduce two novel metrics: *Error L2 Norm (EL2N)* and *Gradient Norm at Initialization (GraNd)*. These metrics aim to provide a more effective means of dataset pruning. It is important to emphasize that the GraNd score at initialization is calculated before any training has taken place, averaging over several randomly initialized models. This fact has been met with skepticism by reviewers[2], but Paul et al. (2021) specifically remark on GraNd at initialization:

> **Pruning at initialization.** In all settings, GraNd scores can be used to select a training subset at initialization that achieves test accuracy significantly better than random, and in some cases, competitive with training on all the data. This is remarkable because GraNd only contains information about the gradient norm at initializion, averaged over initializations.

[1]See https://github.com/google/flax/commit/28fbd95500f4bf2f9924d2560062fa50e919b1a5.
[2]See also https://openreview.net/forum?id=Uj7pF-D-YvT&noteId=qwy3HouKSX.

> This suggests that the geometry of the training distribution induced by a random network contains a surprising amount of information about the structure of the classification problem.

GraNd. The GraNd score measures the magnitude of the gradient vector for a specific input sample in the context of neural network training over different parameter draws. The formula for calculating the (expected) gradient norm is:

$$\text{GraNd}(x) = \mathbb{E}_{\theta_t}[\|\nabla_{\theta_t} L(f(x;\theta_t), y)\|_2] \tag{1}$$

where $\nabla_{\theta_t} L(f(x;\theta_t), y)$ is the gradient of the loss function L with respect to the model's parameters θ_t at epoch t, $f(x;\theta)$ is the model's prediction for input x, and y is the true label for the input. We take an expectation over several training runs. The gradient norm provides information about the model's sensitivity to a particular input and helps in identifying data points that have a strong influence on the learning process.

EL2N. The EL2N score measures the squared difference between the predicted and (one-hot) true labels for a specific input sample. The formula for calculating the EL2N score is:

$$\text{EL2N}(x) = \mathbb{E}_{\theta_t}[\|f(x;\theta_t) - y\|_2^2] \tag{2}$$

where $f(x;\theta)$ is the model's prediction for input x, y is the (one-hot) true label for the input, and $\|\cdot\|_2$ denotes the Euclidean (L2) norm. The EL2N score provides insight into the model's performance on individual data points, allowing for a more targeted analysis of errors and potential improvements.

The GraNd and EL2N scores are proposed in the context of dataset pruning, where the goal is to remove less informative samples from the training data. Thus, one can create a smaller, more efficient dataset that maintains the model's overall performance while reducing training time and computational resources.

While GraNd at initialization does not require model training, it requires a model and is not cheap to compute. In contrast, the input norm of training samples is incredibly cheap to compute and would thus provide an exciting new baseline to use for data pruning experiments. We investigate this correlation in this paper and find positive evidence for it. However, we also find that the GraNd score at initialization does not outperform random pruning, unlike the respective results of Paul et al. (2021) for GraNd at initialization.

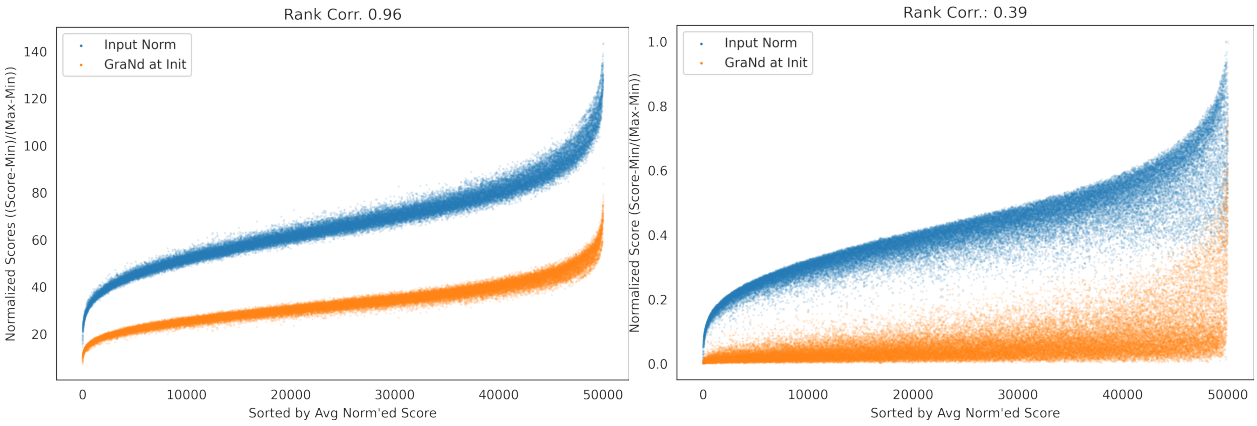
Outline. In §2.1, we begin by discussing the correlation between input norm and gradient norm at initialization. We empirically find strong correlation between GraNd scores at initialization and input norms as we average over models. In §2.2, we explore the implication of this insight for dataset pruning and find that both GraNd at initialization and input norm scores do not outperform random pruning, but GraNd scores after a few epochs perform similar to EL2N scores at these later epochs.

In summary, this reproduction contributes a new insight on the relationship between input norm and gradient norm at initialization and finds a failure to reproduce one of the six contributions of Paul et al. (2021).

2 Investigation

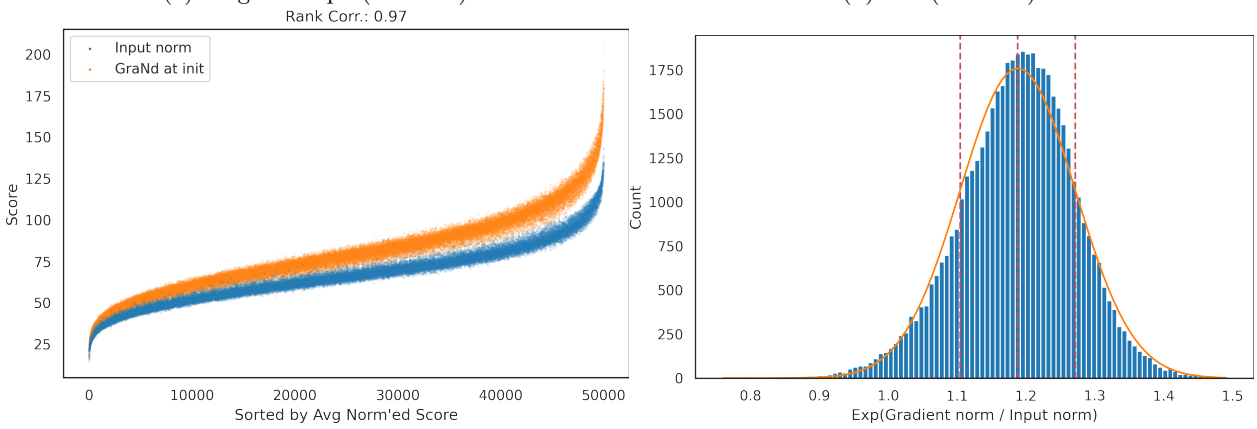
We investigate the correlation between input norm and GraNd at initialization and the other scores on CIFAR-10 (Krizhevsky et al., 2009) in three different ways: First, we update the original paper repository[3] (`https://github.com/mansheej/data_diet`), which uses JAX (Bradbury et al., 2018), rerun the experiments for Figure 1 (second row) in Paul et al. (2021) for CIFAR-10, which trains for 200 epochs, using GraNd at initialization, GraNd at epoch 20, E2LN at epoch 20, Forget Score at epoch 200, and input norm. Second, we reproduce the same experiments using 'hlb' (Balsam, 2023), which is a strongly modified version of ResNet-18 that allows to train to high accuracy in 12 epochs taking about 30 seconds total on a Nvidia RTX 4090 in PyTorch (Paszke et al., 2019) For the latter, we compare GraNd at initialization, GraNd at epoch 1 ($\approx 20/200 \cdot 12$ epochs), EL2N at epoch 1, and input norm[4]. Third, we compare the rank correlations between

[3] `https://github.com/XXX/data_diet`
[4] `https://github.com/XXX/pytorch_datadiet`



(a) Original Repo (10 trials)

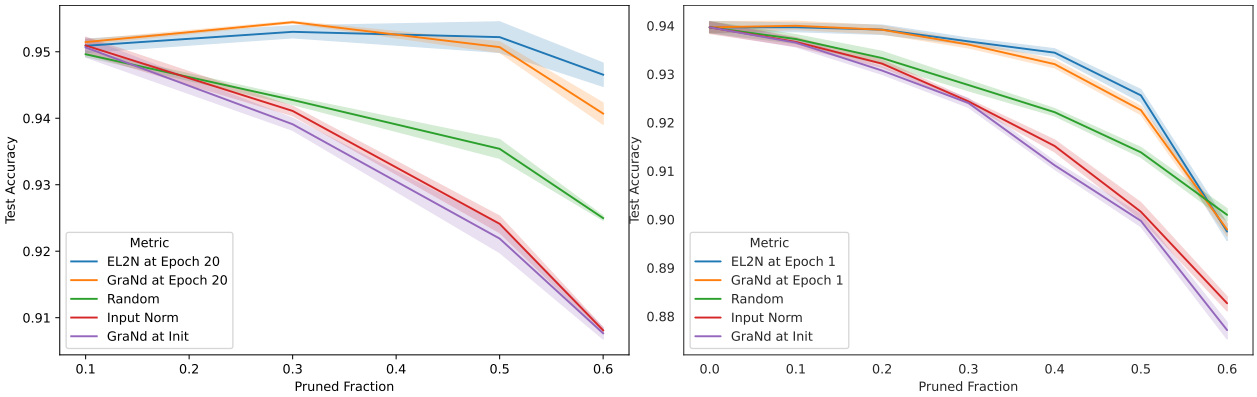
(b) Hlb (10 trials)



(c) Minimal (120 trials)

(d) Minimal (120 trials)

Figure 1: *Correlation between GraNd at Initialization and Input Norm for CIFAR-10's training set.* **(a, b, c)**: We sort the samples by their average normalized score (i.e., the score minus its minimum divided by its range), plot the scores and compute Spearman's rank correlation on CIFAR-10's training data. The original repository and the 'minimal' implementation have very high rank correlation—'hlb' has a lower but still strong rank correlation. **(d)**: *Ratio between input norm and gradient norm.* In the 'minimal' implementation, the ratio between input norm and gradient norm is roughly log-normal distributed.



(a) Original Repo (2 trials each)

(b) Hlb (10 trials)

Figure 2: *Reproduction of Figure 1 (second row) from Paul et al. (2021).* In both reproductions, GraNd at initialization performs as well as the input norm. However, it does not perform better than random pruning. Importantly, it also fails to reproduce the results from Paul et al. (2021). However, GraNd at epoch 20 (respectively at epoch 1 for 'hlb') performs similar to EL2N and like GraNd at initialization in Paul et al. (2021).

the different scores for those two repositories and also use another 'minimal' CIFAR-10 implementation (van Amersfoort, 2021) with a standard ResNet18 architecture for CIFAR-10 to compare the rank correlations.

2.1 Correlation between GraNd at initialization and input norm

To better understand the relationship between the input norm and the gradient norm at initialization, let us consider a toy example first and then appeal to empirical evidence as is common in deep learning research: let's examine linear softmax classification with C classes (without a bias term). The model takes the form:

$$f(x) = \text{softmax}(Wx), \tag{3}$$

together with the cross-entropy loss function:

$$L = -\log f(x)_y. \tag{4}$$

The gradient of the loss function with respect to the rows w_j of the weight matrix W is:

$$\nabla_{w_j} L = (f(x)_j - \mathbb{1}\{j = y\})x \tag{5}$$

where $\mathbb{1}\{j = y\}$ is the indicator function that is 1 if $j = y$ and 0 otherwise. The squared norm of the gradient is:

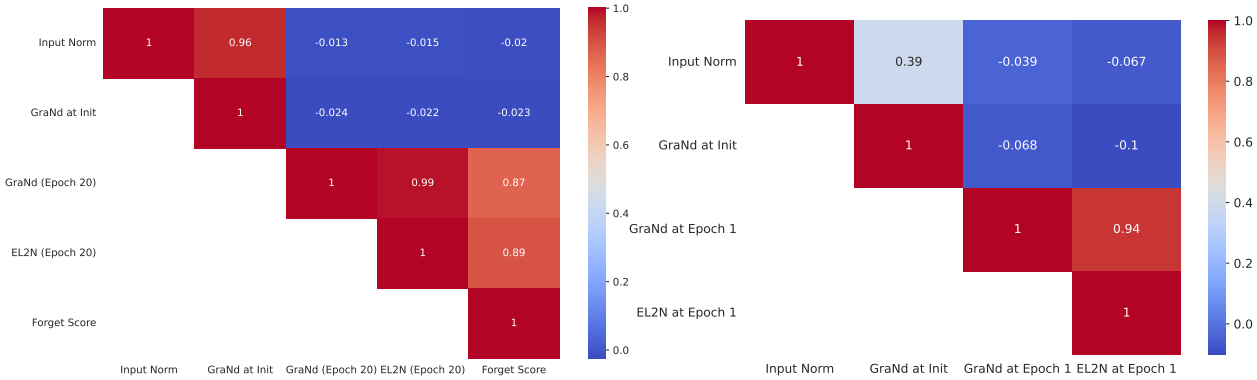
$$\|\nabla_w L\|_2^2 = \sum_{j=1}^{C} (f(x)_j - \mathbb{1}\{j = y\})^2 \|x\|_2^2. \tag{6}$$

In expectation over W (different initializations), the norm of the gradient is:

$$\mathbb{E}_W \left[\|\nabla_w L\|_2\right] = \mathbb{E}_W \left[\left(\sum_{j=1}^{C} (f(x)_j - \mathbb{1}\{j = y\})^2 \right)^{1/2} \right] \|x\|_2. \tag{7}$$

Thus, we see that the gradient norm is a multiple of the input norm. The factor depends on $f(x)_j$, which we could typically expect to be $1/C$.

Empirical Evidence. In Figure 1, we see that on CIFAR-10's training set, GraNd at initialization and the input norm are highly correlated. This is true for the original repository, the 'hlb' and the 'minimal' implementation. The 'hlb' implementation has a lower but still strong correlation.



(a) Original Repo (2 trials each)

(b) Hlb (10 trials)

Figure 3: *Rank Correlations of the Scores.* Cf. Figure 12 in the appendix of Paul et al. (2021). In both reproductions, GraNd at initialization and input norm are positively correlated, while GraNd and EL2N at later epochs are strongly correlated with each other and the Forget Score (at epoch 200).

2.2 Reproducing Figure 1 of Paul et al. (2021) on CIFAR-10

In Figure 2, we see that GraNd at initialization performs about as well as using the input norm. However, it does not reproduce the results from Paul et al. (2021). It performs worse than random pruning (for 'hlb'). However, GraNd at epoch 20 (respectively at epoch 1 for 'hlb') performs like GraNd at initialization in Paul et al. (2021). Similarly, in Figure 3, we see that GraNd at initialization and the input norm are strongly correlated as are GraNd at later epochs, EL2N and the Forget Score, with little correlation between these two groups.

3 Conclusion

If GraNd at initialization performed as well as claimed in Paul et al. (2021), using the input norm would provide a new exciting baseline for data pruning because it is model independent and cheaper to compute than GraNd or other scores. However, since only GraNd at later epochs seems to perform as expected, we cannot recommend using input norm or GraNd at initialization for data pruning.

As to the failure to reproduce the results of Paul et al. (2021), we could not rerun the code using the original JAX version because it is too old for our GPU. The authors of Paul et al. (2021) were, however, able to set up a Google Cloud VM with an old image that was able to reproduce the original results using the original JAX version. On further investigation, the author of this reproduction found a bug in `flax.training.restore_checkpoint` that was fixed in April 2021[5]: passing a `0` step (i.e. initialization) would trigger loading the *latest* checkpoint instead of the zero-th checkpoint because the internal implementation was checking `if step:` instead of `if step is not None:` when deciding whether to fallback to loading the latest checkpoint. This bug was fixed in April 2021, but the authors of Paul et al. (2021) were not aware of this bug and did not rerun their experiments with newer JAX/FLAX versions. We have accordingly informed the authors of Paul et al. (2021).

References

Tysam Balsam. hlb-CIFAR10, 2 2023. URL https://github.com/tysam-code/hlb-CIFAR10.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

[5]See https://github.com/google/flax/commit/28fbd95500f4bf2f9924d2560062fa50e919b1a5.

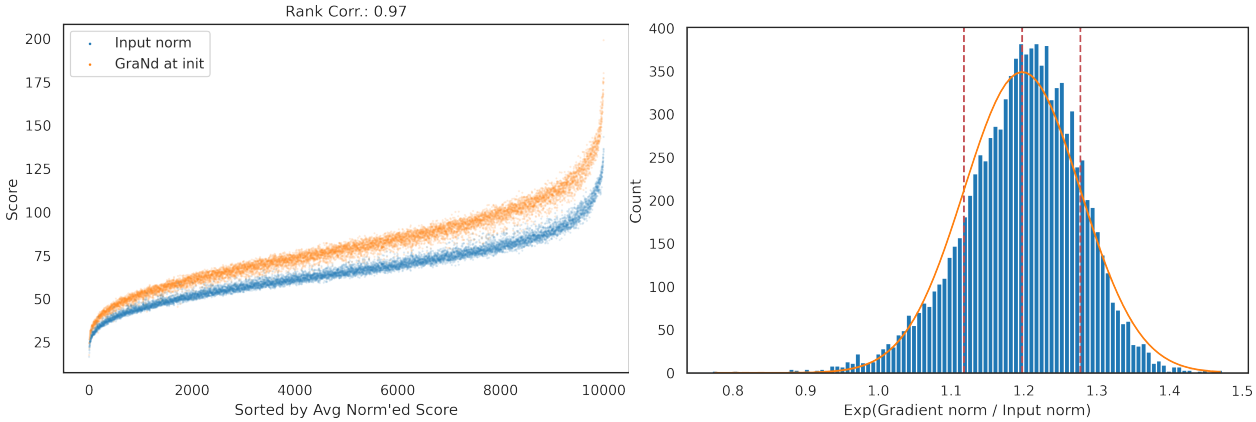
Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607, 2021.

Joost van Amersfoort. Minimal CIFAR-10, 5 2021. URL https://github.com/y0ast/pytorch-snippets/tree/main/minimal_cifar.

A Appendix



(a) Minimal (1000 trials)

(b) Minimal (1000 trials)

Figure 4: *Correlation between GraNd at Initialization and Input Norm on the Test Set.* (a): We sort the samples by their average normalized score (i.e., the score minus its minimum divided by its range), plot the scores and compute Spearman's rank correlation on CIFAR-10's test data. The original repository and the 'minimal' implementation have very high rank correlation—'hlb' has a lower but still strong rank correlation. (b): *Ratio between input norm and gradient norm.* In the 'minimal' implementation, the ratio between input norm and gradient norm is roughly log-normal distributed