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) presents two novel metrics, GraNd (at initialization) and EL2N, for pruning datasets during the training of neural networks. The GraNd score at initialization computes the average gradient norm of an input sample across multiple randomly initialized models before any training has occurred. Our analysis uncovers a strong correlation between the GraNd score at initialization and the input norm of a sample, suggesting that the latter could have served as an inexpensive new baseline for data pruning. Although we successfully replicated the results for the EL2N score at epoch 20, we could not do the same for the GraNd score at initialization, contradicting one of the findings in Paul et al. (2021). Regrettably, neither the GraNd score at initialization nor the input norm outperforms random pruning in terms of performance. However, the GraNd scores later in training do provide useful pruning signals. Overall, we could not reproduce the original CIFAR-10 results for GraNd at initialization using both an updated version of the original JAX repository and a newly implemented PyTorch codebase. Upon investigating the underlying JAX/FLAX code from 2021, we discovered a bug in the checkpoint restoring code that was resolved in April 2021 as the cause of the initially reported results[1].

1 Introduction

Deep learning has led to significant advances in various applications, such as computer vision, natural language processing, and speech recognition. However, the vast amounts of data required to train neural networks pose challenges in computational resources, storage, and energy consumption. Consequently, there is a growing interest in developing methods to reduce data requirements while maintaining model performance.

The authors of 'Deep Learning on a Data Diet' (Paul et al., 2021) introduced novel metrics for pruning datasets during the training of neural networks. Recently, the senior author presented their work in a talk, highlighting the surprising effectiveness of GraNd at initialization, which prompted the author of this paper to ask about *a potential correlation between the proposed GraNd score at initialization and input norms*. The input norm is cheap to compute and could thus provide an intriguing new baseline for data pruning experiments. This research question led to further investigation. In this paper, we explore 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

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

contains information about the gradient norm at initializion, averaged over initializations. 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.

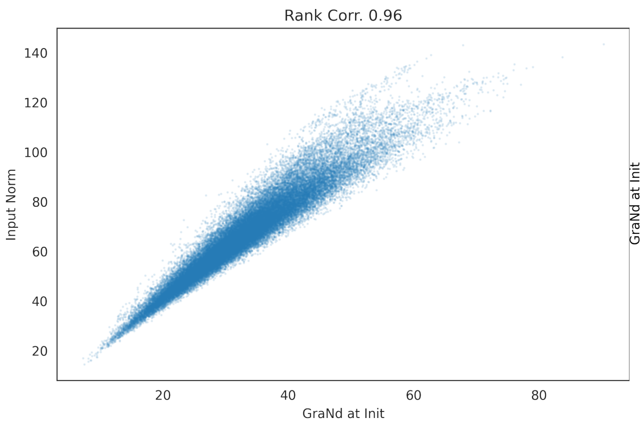
In summary, this reproduction contributes a new insight on the relationship between input norm and gradient norm at initialization and finds and fully explains for a failure to reproduce one of the six contributions of Paul et al. (2021).

2 Investigation

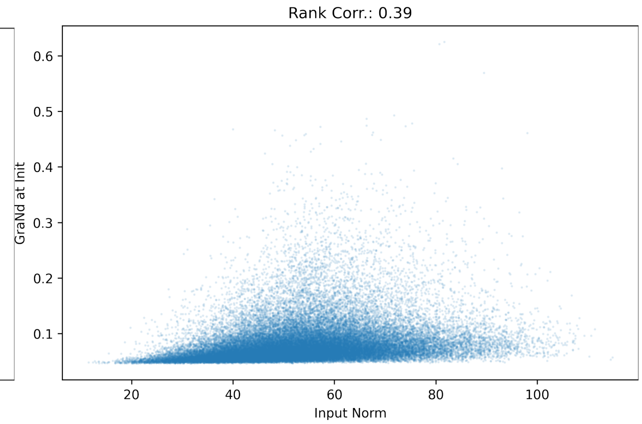
We examine the correlation between the input norm and GraNd at initialization, as well as other scores on CIFAR-10 (Krizhevsky et al., 2009), through three distinct approaches:

1. First, we update the original paper repository[3] (`https://github.com/mansheej/data_diet`), which utilizes JAX (Bradbury et al., 2018). We rerun the experiments for Figure 1 (second row) in Paul et al. (2021) on CIFAR-10, training for 200 epochs using GraNd at initialization, GraNd at epoch 20, E2LN at epoch 20, Forget Score at epoch 200, and input norm.

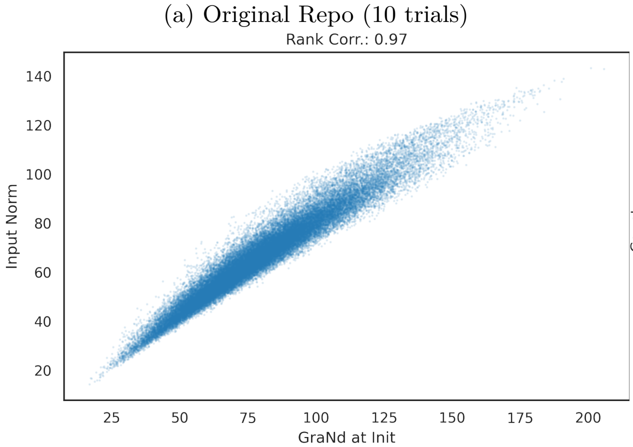
[3]See `https://github.com/xxx/data_diet`



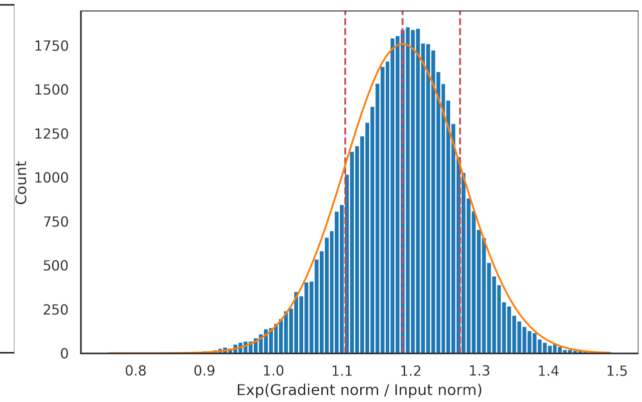
(a) Original Repo (10 trials)



(b) Hlb (10 trials)

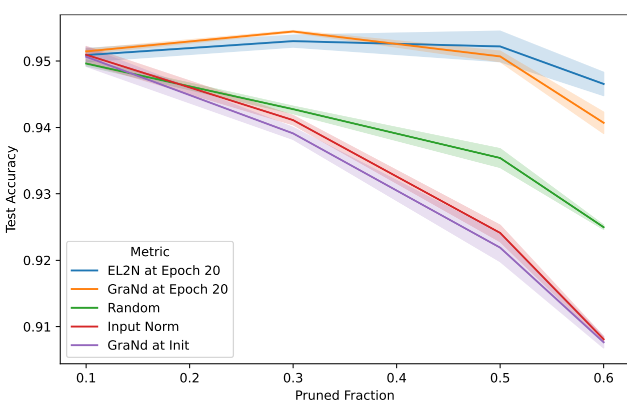


(c) Minimal (120 trials)

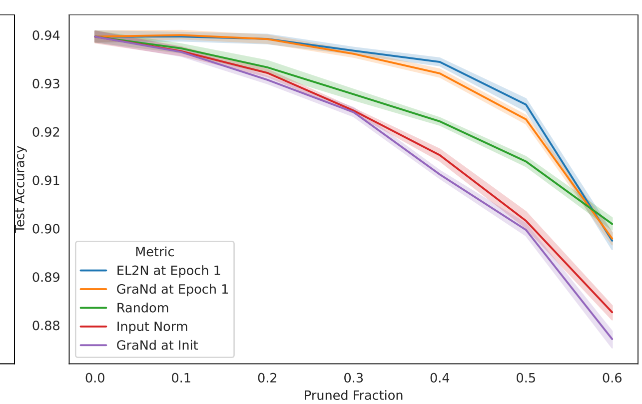


(d) Minimal (120 trials)

Figure 1: *Correlation between GraNd at Initialization and Input Norm for CIFAR-10's training set.* (**a**, **b**, **c**): 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).

2. Second, we reproduce the same experiments using 'hlb' (Balsam, 2023), a significantly modified version of ResNet-18 that achieves high accuracy in 12 epochs, taking approximately 30 seconds in total on an Nvidia RTX 4090 with PyTorch (Paszke et al., 2019). For this approach, we compare GraNd at initialization, GraNd at epoch 1 ($\approx 20/200 \cdot 12$ epochs), EL2N at epoch 1, and input norm[4]. We analyze the rank correlations between the different scores for the two repositories mentioned above.

3. Third, we employ 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 gain a deeper understanding of the relationship between the input norm and the gradient norm at initialization, we first consider a toy example and then provide empirical evidence. Let us examine a linear softmax classifier with C classes (without a bias term). The model takes the form:

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

together with the cross-entropy loss function:

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

The gradient of the loss function concerning 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:

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

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

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

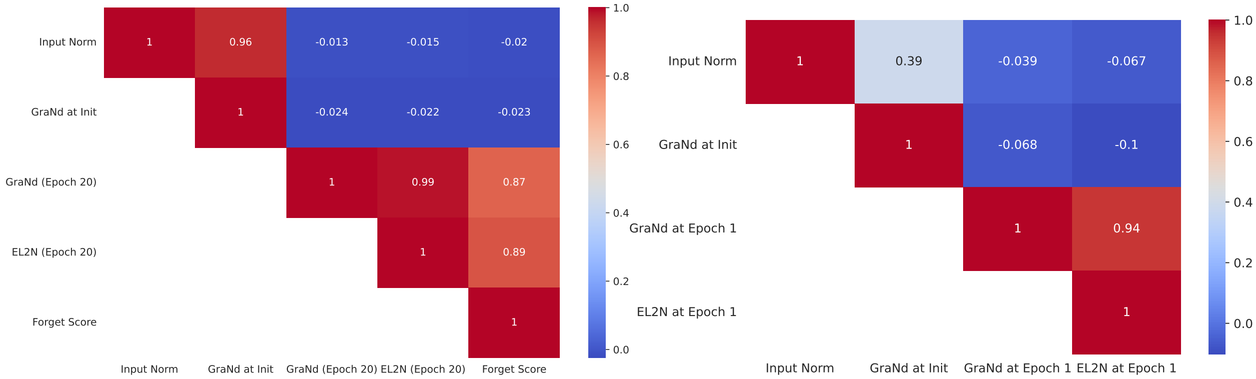
As a result, the gradient norm is a multiple of the input norm. The factor depends on $f(x)_j$, which we could typically expect to be $1/C$.

Empirical Evidence. In Figure 1, we observe that on CIFAR-10's training set, GraNd at initialization and the input norm exhibit a high correlation. This holds true for the original repository, the 'hlb', and the 'minimal' implementations. The 'hlb' implementation demonstrates a lower, yet still evident, correlation.

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

Figure 2 demonstrates that GraNd at initialization performs similarly to using the input norm. However, it does not reproduce the results from Paul et al. (2021). It performs worse than random pruning (for 'hlb'). Nonetheless, GraNd at epoch 20 (respectively at epoch 1 for 'hlb') exhibits performance akin to GraNd at initialization in Paul et al. (2021). Additionally, Figure 3 reveals that GraNd at initialization and the input norm are strongly correlated, as are GraNd at later epochs, EL2N, and the Forget Score, with minimal correlation between these two groups.

[4]https://github.com/xxx/pytorch_datadiet



(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).

3 Conclusion

If GraNd at initialization had performed as well as claimed in Paul et al. (2021), using the input norm could have provided a new and exciting baseline for data pruning, given its model independence and lower computational cost compared to GraNd or other scores. However, since only GraNd at later epochs appears to perform as expected, we cannot recommend using the input norm or GraNd at initialization for data pruning.

Regarding the failure to reproduce the results of Paul et al. (2021), we were unable to rerun the code using the original JAX version due to its incompatibility with our GPU. Nevertheless, the authors of Paul et al. (2021) managed to set up a Google Cloud VM with an older image that could reproduce the original results using the original JAX version. Upon further investigation, the author of this reproduction discovered 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 it and did not rerun their experiments with newer JAX/FLAX versions. We have duly 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.

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

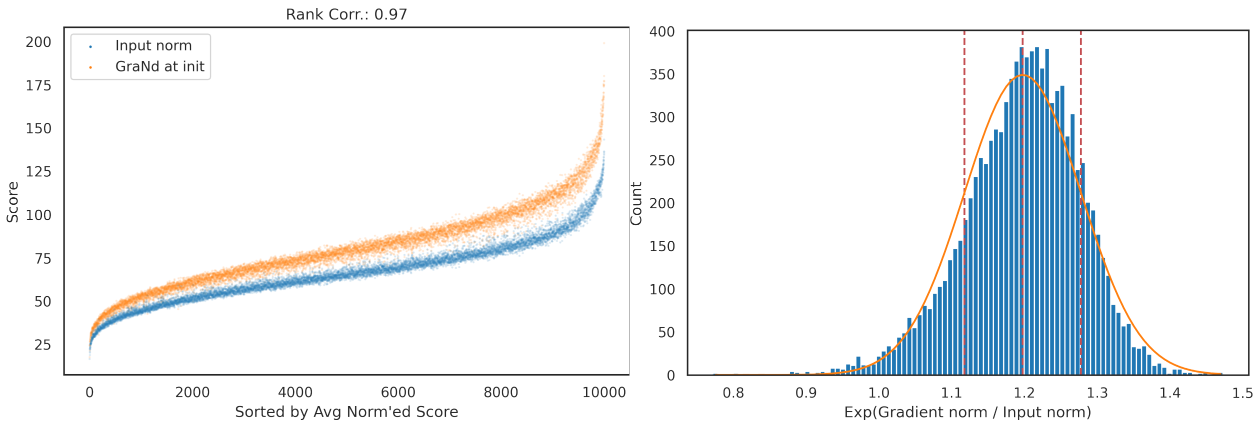
Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith 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.

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

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