A Short Empirical Note on Scaling Behavior in Small Neural Networks

Ritvik Chappidi
Mathematics
Rock Ridge High School
Ashburn, Virginia
chappidi.ritvik@gmail.com

Aditya Jupally
Computer Science
Rock Ridge High School
Ashburn, Virginia
adityajupally@gmail.com

***Abstract*—Scaling laws describe how model performance improves with dataset size, model width, and compute. While such laws are well documented for large-scale language models, their behavior in small networks remains less understood. This paper presents a concise empirical study of loss scaling behavior in simple feedforward neural networks trained on synthetic regression tasks. Results show that even very small networks follow an approximate power-law relationship between dataset size and test loss, with a fitted exponent of about 0.076. These findings suggest that scaling regularities emerge even at small scales, implying that the underlying principles of efficiency and generalization extend beyond large-scale models.**

***Index Terms*—Scaling laws, neural networks, statistical learning, power-law, machine learning theory**

I. Introduction

Recent research has identified power-law relationships between model scale and performance across deep learning domains [1], [3]. The empirical law $L \propto N^{-\alpha}$, where L denotes test loss and N the dataset size, captures the diminishing returns of larger datasets. However, existing analyses are primarily confined to billion-parameter regimes. This work investigates whether analogous scaling trends emerge in small networks with limited capacity.

The motivation is twofold. First, verifying the existence of scaling behavior at micro-scales would support the universality of these empirical laws. Second, understanding scaling in small models provides a cost-efficient method for predicting model performance without large-scale computation.

II. Methods

We trained a single-hidden-layer multilayer perceptron (MLP) using the `scikit-learn` `MLPRegressor`. The regression target was generated as:

$$y = \sum_{i=1}^{5} \sin(x_i) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.1) \tag{1}$$

where $x_i \in [-1, 1]$. Dataset sizes $N \in \{100, 500, 1000, 5000, 10000\}$ were used. For each N, models were trained for 500 iterations with 64 hidden units using the Adam optimizer and mean-squared error loss.

The empirical scaling relationship is modeled as:

$$L = aN^{-\alpha} + b \tag{2}$$

where α represents the scaling exponent. Parameters (a, b, α) were estimated via nonlinear least squares on $\log L$ versus $\log N$.

III. Results and Discussion

The observed test loss decreased smoothly as dataset size increased, with diminishing returns for large N. The fitted scaling law yielded $\alpha = 0.076$ (Fig. 1), indicating that even small neural networks exhibit measurable scaling patterns. The nearly linear trend in log-log space supports the hypothesis that performance scaling is a fundamental property of stochastic optimization rather than an artifact of large-scale training.

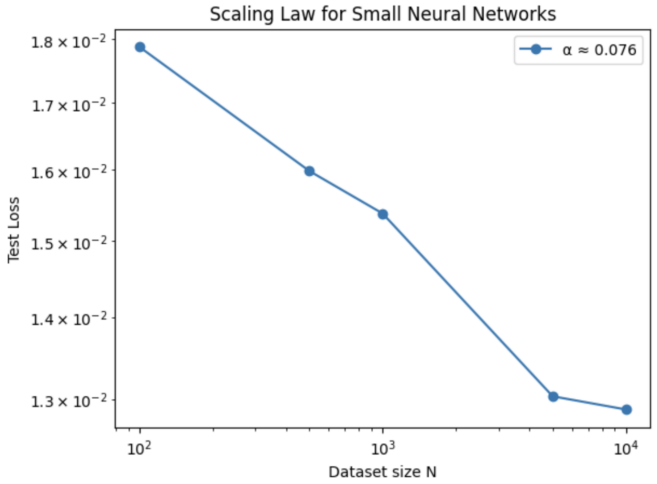

Fig. 1. Log-log plot of test loss versus dataset size N for small neural networks. The fitted power-law exponent is $\alpha \approx 0.076$.

While the scaling exponent is smaller than that observed in large models [2], the persistence of this pattern across three orders of magnitude in N demonstrates that small-scale models share the same statistical structure. These results imply that researchers can estimate scaling exponents with minimal computation before committing to large-scale experiments.

IV. Conclusion

This work demonstrates that even modest neural networks trained on synthetic data follow approximate scaling laws between dataset size and loss. The recovered exponent $\alpha \approx 0.076$

aligns qualitatively with literature on large models, reinforcing the idea that statistical scaling may emerge from generalization behavior intrinsic to stochastic gradient descent. Future research should extend this framework to small convolutional or recurrent networks and explore connections between α, regularization, and architectural depth.

REFERENCES

[1] J. Kaplan et al., "Scaling laws for neural language models," arXiv preprint arXiv:2001.08361, 2020.

[2] J. Hoffmann et al., "Training compute-optimal large language models," arXiv preprint arXiv:2203.15556, 2022.

[3] T. Henighan et al., "Scaling laws for autoregressive generative modeling," arXiv preprint arXiv:2301.11565, 2023.