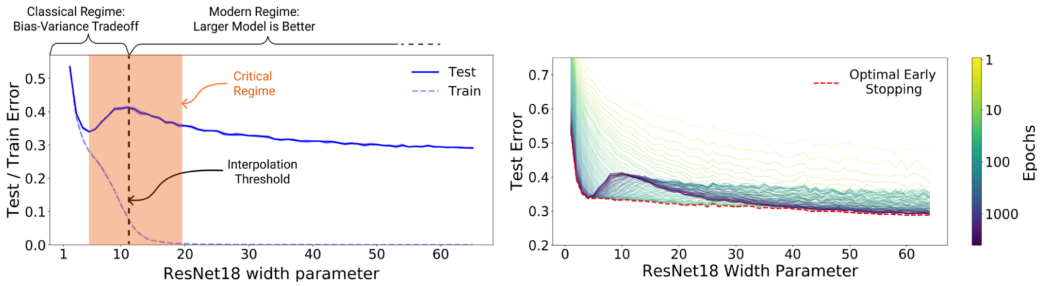
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Project: Sample Efficiency in Text Diffusion Pretraining



Credit to Nakkiran et al. [6]: The Deep double descent phenomemon : given a fixed dataset size, a the test performance of a model will plateau as it is scaled past overfitting

Despite being invented in 2015 [7], diffusion image models only came into prominence in 2020. Despite being more computationally intensive to train, such models are more *sample efficient*: given a fixed dataset size they can learn to generate much higher quality samples than their alternatives, provided they have sufficiently large computational resources for training.

Recently, diffusion models have been applied [1] and scaled on text data, and seem poised to challenge autoregressive (predict-the-next-token) LLMs for state-of-the-art text generation. The question we would like to answer with this project is: "Is diffusion pretraining of LLMs more sample efficient autoregressive pretraining?".

In the current year of 2025, this question is particularly salient: frontier LLMs are transitioning from a regime where compute is scarce to a regime where data are scarce. Following the so-called scaling laws, which determine the most compute-efficient dataset sizes to train autoregressive LLMs, frontier models are now trained on all ≈ 15 Trillion useful tokens on the internet [3], with Ilya Sutskever referring to text data as the "fossil fuel of AI". Improving sample efficiency also has implications for capabilities: currently, autoregressive LLMs struggle to generalize past their vast training dataset, which covers $> 10,000$x more language input than and single person receives in their lifetime. If diffusion models are more sample efficient than their alternatives for text generation (as they are for image generation), this could mean that they will exhibit greater capabilities once there exists sufficient computational resources to scale them to these large sizes.

Practically, this project will look like exploring & replicating results such as the "Kaplan" [5] and "chinchilla" [4] scaling laws and deep double descent [2, 6] with a small dataset for both diffusion and autoregressive LLMs to investigate their sample efficiency.

Primary supervisor

- Sam Dauncey: sdauncey@ethz.ch, ETZ G61.1

References

[1] Jacob Austin et al. Structured denoising diffusion models in discrete state-spaces. In: *Advances in neural information processing systems* 34 (2021), pp. 17981–17993.

[2] Mikhail Belkin et al. Reconciling modern machine-learning practice and the classical bias–variance trade-off. In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854.

[3] Aaron Grattafiori et al. The Llama 3 Herd of Models. 2024. arXiv: 2407.21783 [cs.AI].

[4] Jordan Hoffmann et al. Training Compute-Optimal Large Language Models. 2022. arXiv: 2203.15556 [cs.CL].

[5] Jared Kaplan et al. Scaling Laws for Neural Language Models. 2020. arXiv: 2001.08361 [cs.LG].

[6] Preetum Nakkiran et al. Deep double descent: Where bigger models and more data hurt. In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.12 (2021), p. 124003.

[7] Jascha Sohl-Dickstein et al. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In: Proceedings of the 32nd International Conference on Machine Learning. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 2256–2265.