Stat 9911
Principles of AI: LLMs
Key Empirical Behaviors of LLMs

Edgar Dobriban

Department of Statistics and Data Science, the Wharton School, University of Pennsylvania

February 14, 2025

Plan

- We plan to discuss some key empirical behaviors of LLMs.

Table of Contents

Scaling Laws for LLMs

- Scaling laws are empirical observations about the test loss of LLMs.

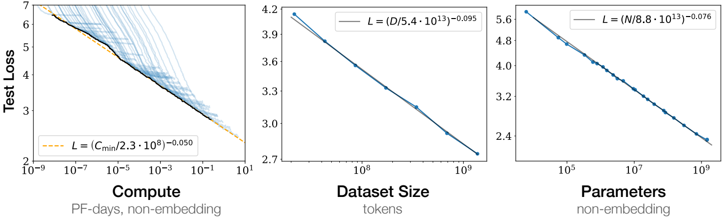

Figure: Kaplan et al. (2020)

- Let D be the training dataset size (# tokens) and N be the number of non-embedding parameters in an LLM.
- Let $L(\cdot)$ denote the test perplexity achieved by an LLM (or, the best among a few possibilities).

Parameter Count for Transformer

- For each layer:
  - For each head:
    - Queries, Keys, Values: W_q, W_k, W_v, each $d' \times d$, where d is embedding dim, and d' is attention dim. Total $3Hdd'$
    - Output projection W_o is $Hd' \times d$. Total Hdd'
  - FFN: W_1 is $d_{ff} \times d$, W_{proj} is $d \times d_{ff}$. Total $2dd_{ff}$.
  - Total per layer: $N_1 = 4Hdd' + 2dd_{ff}$. Often $d' = d/H$, $d_{ff} = 4d$, so $N_1 = 4d^2 + 8d^2 = 12d^2$
- Overall $N = N_1 n_{\mathrm{layer}} = 12 n_{\mathrm{layer}} d^2$
- Exclude initial token embeddings, positional encoding

Kaplan et al. (2020) Scaling Law

- Kaplan et al. (2020) found that for some scalars $\alpha_N, \alpha_D > 0$, $N_c, D_c > 0$,

$$L(N, D) \approx \left[\left(\frac{N_c}{N} \right)^{\alpha_N/\alpha_D} + \left(\frac{D_c}{D} \right) \right]^{\alpha_D}$$

- N_c, D_c: Critical values above which scaling laws hold.
- Holds over several orders of magnitude of N, D.
- Performance decreases as a power law:

$$L(N, D) \sim \frac{1}{N^{\alpha_N}} + \frac{1}{D^{\alpha_D}}.$$

- They find $\alpha_N \approx 0.076$, $\alpha_D \approx 0.095$

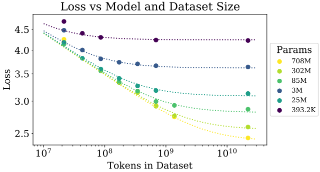

Figure: Kaplan et al. (2020)

Compute for a Transformer

- A $a \times b$ into $b \times c$ matrix-matrix multiplication takes roughly $2abc$ flops (abc multiplications and $a(b-1)c$ additions)
- So in a forward pass, the dominating number of flops is $F_1 = 2N$
- Backward pass/back-propagation: $F_2 \approx 2F_1$
  - Simplest to see this for a matrix operation $y = Wx$, where x is d-dim, W is $d \times d$
  - Forward pass $\approx 2d^2$ flops.
  - Backward pass: Compute $\frac{\partial L}{\partial x} = W^\top \cdot \frac{\partial \mathcal{L}}{\partial y}$, where $\frac{\partial \mathcal{L}}{\partial y}$ is $d \times 1$ [total $2d^2$]
  - Then $W = W - \eta \frac{\partial \mathcal{L}}{\partial W}$, where $\frac{\partial \mathcal{L}}{\partial W} = \frac{\partial \mathcal{L}}{\partial y} \cdot x^\top$ [total $2d^2$]
  - Overall $4d^2$
- Total $6N$; and this is for every token, so $C = 6ND$.
- Exclude positional encoding computation and lower-order terms (biases in FFNs)

Kaplan et al. (2020): Optimal Scaling

- Total compute: $C = 6ND$.
- Given a specific compute budget $C_{\max}$, solve:

$$\min_{N,D} L(N, D) \quad \text{subject to } 6ND \leq C_{\max}.$$

- Optimum: $N^{\alpha_N} \sim D^{\alpha_D}$.
- Example: for $\alpha_N \approx 0.076$, $\alpha_D \approx 0.095$, $D \approx N^{0.8}$, so increase dataset size sublinearly with parameters[1].
- If we consider that $D = BS$, where B is the batch size and S is the number of gradient steps, then, for a given batch size, we can obtain the needed S

[1] Kaplan et al. (2020) write $N^{0.74}$.

Chinchilla Scaling Law (Hoffman et al., 2023)

- Hoffman et al. (2023) found a slightly different scaling law:

$$L(N, D) = \mathcal{E} + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}},$$

  where $\mathcal{E} = 1.69$, $\alpha \approx 0.34$, $\beta \approx 0.28$.
- Suggests roughly equal scaling of model and dataset sizes.

Experimental Validation by Hoffman et al.

- Train models of various architectures, sizes, and dataset sizes.
- Plot smoothed train loss as a function of FLOPs.
- Find lower envelope to validate scaling law.

Decomposition of Loss (Hoffman et al., 2023)

- Decomposition:

$$L(N, D) = L(\hat{f}_{N,D}) = L(f^*) + \big(L(f_N) - L(f^*)\big) + \big(L(\hat{f}_{N,D}) - L(f_N)\big),$$

- L: Population-level risk function.
- $L(f^*)$: Bayes risk.
- $L(f_N) - L(f^*)$: Approximation error for the best model of size N.
- $L(\hat{f}_{N,D}) - L(f_N)$: Random error of the fitted model.

Table of Contents

Emergence (Wei et al., 2022)

- Emergence in general: Quantitative change leads to qualitative change (e.g., uranium, DNA, water).
- For ML: Small models cannot solve a task, but large models can.
- Related concept: Grokking (similar meaning).

Table of Contents

Memorization in LLMs

- LLMs can memorize text.
- **Desirable**: Memorize facts (e.g., "Who was George Washington?").
- **Undesirable**: Memorizing entire novels (e.g., "Harry Potter") due to copyright concerns.
- Detection: Large likelihood ratio $p(x)/p'(x)$, a.k.a perplexity filter (Carlini et al., 2021).

Extractable Memorization (Nasr et al., 2023)

Definition 1: Extractable Memorization

- Given a generation routine Gen, an example x is extractably memorized if an adversary can construct a prompt p such that $\text{Gen}(p) = x$.

Definition 2: Discoverable Memorization

- x is discoverably memorized if $\text{Gen}(p) = x$ when sampling $[p \mid x]$ from the training data.

Prior work: About 1% of training data is discoverably memorized in many LLMs.

Memorization Scores (Biderman et al., 2024)

- **Memorization Score:** For string $S = (S_1, \ldots, S_m)$, start index k, length l, it is the fraction of tokens from $k+1$ to $k+l$ generated by an LLM with prompt $S_{1:k}$ that agree with S.

Memorization and double descent

Table of Contents

Super-Phenomena in LLMs

- **Super-activations (or massive activations)** (Sun et al., 2024):
  - Large activations in specific tokens/dimensions.
  - Values are nearly input-independent.
  - Setting to zero destroys model performance.
- Related to attention sinks (Xiao et al., 2024).

Super-Weights (Yu et al., 2024)

- Super-activations are partly caused by very large weights.
- Modifying them degrades performance completely (e.g., gibberish output).
- In Llama-7B: A single super-weight is more important than the top 7,000 largest weights combined.
- Can be identified using forward passes and examining $e_i' = W_{\text{proj}} \tilde{e}_i$, where $\tilde{e}_i = \sigma(W_1 e_i)$.

Historical Context: Outlier Dimensions

- Earlier work on BERT-busters: Outlier dimensions that disrupt transformers (Kovaleva et al., 2021).
- Similar principles extend to super-phenomena in LLMs.

References

S. Biderman, U. Prashanth, L. Sutawika, H. Schoelkopf, Q. Anthony, S. Purohit, and E. Raff. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium*, pages 2633–2650, 2021.

M. D. Hoffman, D. Phan, david dohan, S. Douglas, T. A. Le, A. T. Parisi, P. Sountsov, C. Sutton, S. Vikram, and R. A. Saurous. Training chain-of-thought via latent-variable inference. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=a147pIS2Co.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

O. Kovaleva, S. Kulshreshtha, A. Rogers, and A. Rumshisky. BERT busters: Outlier dimensions that disrupt transformers. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3392–3405, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.300. URL https://aclanthology.org/2021.findings-acl.300/.

M. Nasr, N. Carlini, J. Hayase, M. Jagielski, A. F. Cooper, D. Ippolito, C. A. Choquette-Choo, E. Wallace, F. Tramèr, and K. Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

References

M. Sun, X. Chen, J. Z. Kolter, and Z. Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

M. Yu, D. Wang, Q. Shan, and A. Wan. The super weight in large language models. *arXiv preprint arXiv:2411.07191*, 2024.