Explaining Neural Scaling Laws

Authors: *Y. Bahri, E. Dyer, J. Kaplan, J. Lee, U. Sharma*
(Google & JHU)

REFORM Reading Group, Oct 29, 2024

Recap: Neural Scaling Laws

- Practical Deepnets are obscenely large and complicated systems.
- Want: predict performance in terms of *available resources*.
- Last week: Empirical evidence that loss on task (e.g., next token prediction) exhibits *power law decay* in [resource type].
- Approach taken: basically fit parametric power law to tons of experiments...
- This session: examine *stylized models* where such behavior arises and is provably quantifiable.

Moving Parts, or: What affects the loss?

- D : size of data set
- P : Number of model parameters
  Initially in paper: exclusively feed-forward NNs of moderate (fixed) depth. P increased by increasing *width*. Note: $W \propto \sqrt{P}$.
- Properties of the data distribution. If data has *intrinsic low-dimensional structure*, expect (hope!) this helps learning.
- Properties of the loss function.
  Not in this presentation.
  Paper gives examples for some pathological cases.
- **Goal:** Scaling laws,

$$\mathcal{L} \propto D^{-\alpha_D}, \qquad \mathcal{L} \propto P^{-\alpha_P} \quad (\text{eqv., } \mathcal{L} \propto W^{-\alpha_W}),$$

  (under different mutual scaling regimes, TBD.)

The Variance-Limited Regime

- Paradigm: **Fix** either D or P; examine loss as other parameter $\to \infty$.

- For this presentation: assume loss is "nice".
  I.e.: twice-differentiable with bounded second derivative.

Variance-limited regime - Dataset scaling

- Fixed P with $D \to \infty$; formally: data is i.i.d. $x_1, \ldots, x_D \sim \mathcal{D}$.
  Claim:
  $$\mathcal{L}(D) \propto D^{-1} + \mathrm{const} \quad \mathrm{as} \quad D \to \infty\,.$$
- *Sketch:*

Variance-limited regime - Dataset scaling

- Fixed P with $D \to \infty$; formally: data is i.i.d. $x_1, \ldots, x_D \sim \mathcal{D}$.
  Claim:
  $$\mathcal{L}(D) \propto D^{-1} + \mathrm{const} \quad \text{as} \quad D \to \infty\,.$$
- *Sketch:*
  - Step I: Let $\hat{f}(x|X_{1:D})$ be the trained model; let $f(x) = \mathbb{E}_{\mathcal{D}}[\hat{f}(x|X_{1:D})]$ be its expectation over data (and possibly randomness in training process). As $D \to \infty$,
    $$\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}[(\hat{f}(x) - f(x))^2] \propto D^{-1}\,.$$
    Intuition: Parametric statistics; nice model (non-singular fisher information) implies MSE $\propto 1/D$.

Variance-limited regime - Dataset scaling

- Fixed P with $D \to \infty$; formally: data is i.i.d. $x_1, \ldots, x_D \sim \mathcal{D}$.
  Claim:
  $$\mathcal{L}(D) \propto D^{-1} + \mathrm{const} \quad \text{as} \quad D \to \infty\,.$$

- *Sketch:*

  - Step I: Let $\hat{f}(x|X_{1:D})$ be the trained model; let $f(x) = \mathbb{E}_{\mathcal{D}}[\hat{f}(x|X_{1:D})]$ be its expectation over data (and possibly randomness in training process). As $D \to \infty$,
    $$\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}[(\hat{f}(x) - f(x))^2] \propto D^{-1}\,.$$
    Intuition: Parametric statistics; nice model (non-singular fisher information) implies MSE $\propto 1/D$.

  - Step II: Taylor-expand the loss around f,
    $$\mathcal{L}[\hat{f}] := \mathbb{E}_D \ell(\hat{f}(x))$$
    $$= \underbrace{\mathbb{E}_{\mathcal{D}}[\ell(f(x))]}_{const} + \underbrace{\mathbb{E}_{\mathcal{D}}[\ell'(f(x))(\hat{f}(x) - f(x))]}_{=0} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\ell''(\xi)(f(x) - \hat{f}(x))^2\right]}_{\propto D^{-1}}.$$

Variance-limited regime - Large width scaling

- Fixed D with $W \to \infty$; random initial weights.
  Claim:
  $$\mathcal{L}(D) \propto W^{-1} + \text{const} \quad \text{as} \quad W \to \infty\,.$$

Variance-limited regime - Large width scaling

- Fixed D with $W \to \infty$; random initial weights.
  Claim:
  $$\mathcal{L}(D) \propto W^{-1} + \text{const} \quad \text{as} \quad W \to \infty \,.$$

- Idea: Previous works shown that $\hat{f}(x) = \hat{f}(x|\Theta)$ converges to a Gaussian process. In particular,
  $$\mathbb{E}_{\Theta}[(\hat{f}(x) - f(x))^2] \propto 1/W,$$
  where $f(x) := \mathbb{E}_{\Theta}[\hat{f}(x|\Theta)]$. Same reasoning as before.

Variance-limited regime - Large width scaling

- Fixed D with $W \to \infty$; random initial weights.
  Claim:
  $$\mathcal{L}(D) \propto W^{-1} + \text{const} \quad \text{as} \quad W \to \infty\,.$$

- Idea: Previous works shown that $\hat{f}(x) = \hat{f}(x|\Theta)$ converges to a Gaussian process. In particular,
  $$\mathbb{E}_{\Theta}[(\hat{f}(x) - f(x))^2] \propto 1/W,$$
  where $f(x) := \mathbb{E}_{\Theta}[\hat{f}(x|\Theta)]$. Same reasoning as before.

- **Summary:** Exponents in variance limited regime,
  $$\alpha_D = \alpha_W = 1\,.$$
  Universal (under "niceness" assumption).

Resolution-limited regime

- Previously we fixed either D or P. Rather boring scaling.
- Now: *both* $P, D \gg 1$. Two cases:
  (1) Over-parameterized: $P \gg D \gg 1$,
  (2) Under-parameterized: $D \gg P \gg 1$.
- Furthermore, assume data has intrinsic low dimension.
  Specifically, $x_1, \ldots, x_D$ are i.i.d. uniform on low-dim compact manifold $\mathcal{M}_d$.
  Corresponding labels: $y_i = f(x_i)$, $f : \mathcal{M}_d \to \mathbb{R}$.

Warm up I: Over-parameterized regime

- Suppose $P \gg D$. In this regime, network has enough parameters to *memorize* the data.
  Assume $\hat{f}$ *interpolates* data, e.g. zero training loss: $\hat{f}(x_i) = f(x_i)$ for all $1 \leq i \leq d$.

- For $x \in \mathcal{M}_d$, let $\hat{x}_{NN}(x)$ be its nearest neighbor among $x_1, \ldots, x_D$.
$$|x - \hat{x}_{NN}(x)| \asymp D^{-1/d}, \qquad x, x_1, \ldots, x_D \sim \mathcal{M}_d.$$

- For an interpolator,
$$\begin{aligned} |f(x) - \hat{f}(x)| &= \left| f(x) - f(\hat{x}_{NN}) + f(\hat{x}_{NN}) - \hat{f}(x) \right| \\ &= \left| f(x) - f(\hat{x}_{NN}) + \hat{f}(\hat{x}_{NN}) - \hat{f}(x) \right| \\ &\leq |f(x) - f(\hat{x}_{NN})| + |\hat{f}(\hat{x}_{NN}) - \hat{f}(x)| \\ &\leq (\|f\|_{Lip} + \|\hat{f}\|_{Lip})|x - \hat{x}_{NN}|. \end{aligned}$$
  Therefore, assuming $f, \hat{f}$ have bounded Lipshitz constants,
$$\mathcal{L}[\hat{f}] := \mathbb{E}_{x, x_{1:D}}[|f(x) - \hat{f}(x)|] \lesssim D^{-1/d}.$$
  (Theorem 2 in paper.)

Warm up II: Over-parameterized regime

- Suppose $D \gg P$.
  No capacity (parameters) to memorize entirely.
  Enough parameters to memorize $O(P)$ data points and their labels.

- Let $\hat{f}$ be the rule that interpolates on $O(P)$ random pairs (x_i, y_i).
  By previous argument,

$$\mathcal{L}[\hat{f}] \lesssim P^{-1/d} .$$

(Theorem 3 in paper.)

Why are you telling me this?

- Implicit assumption: NNs can do better than simple interpolators.
- These are *upper bounds*; not necessarily saturated in practice.
- Take home message: scaling law potentially depends on data.
  Should $\mathcal{L}(D) \propto D^{-c/d}, P^{-c/d}$ for some (meaningful) $c > 0$?
- Curiously (well, by design), data and parameters exponent upper bound are the same.

The Random Features Model

- Let $\{F_i : \mathcal{M}_d \to \mathbb{R}\}_{i=1}^S$, be a collection of *features*. Denote

$$\boldsymbol{F}(x) = [F_1(x), \cdots, F_S(x)] : \mathcal{M}_d \to \mathbb{R}^S .$$

  I.e.: random feature mappings; last layer of trained NN, NTK...
  Here $S \gg D, P$.

- Motivation: Previous work has shown that real-world NNs can be approx'd—to a degree—by suitable feature models.
  I.e.: the Neural Tangent Kernel (NTK). Features correspond to linearization of $\hat{f}$ around the initial (random) weights.

- The paper considers a teacher-student model, as follows.
- Teacher:

$$y_i = \boldsymbol{\omega}^\top \boldsymbol{F}(x_i), \qquad x_i \sim \mathcal{M}_d, \qquad 1 \le i \le D.$$

  Isotropic prior $\omega \sim N(0, S^{-1} I_S)$.

- Student: uses P "features". Fix matrix $\mathcal{P} \in \mathbb{R}^{P \times S}$, P is available # of features.
  E.g., $\mathcal{P}$ choose P random features uniformly.
  The student features are $\mathbf{f}(x) := \mathcal{P}\boldsymbol{F}(x) : \mathcal{M}_d \to \mathbb{R}^P$.

- Denote

$$\underline{\boldsymbol{F}} = [\boldsymbol{F}(x_1); \cdots ; \boldsymbol{F}(x_D)]^\top \in \mathbb{R}^{D \times S},$$
$$\underline{\boldsymbol{f}} = [\boldsymbol{f}(x_1); \cdots ; \boldsymbol{f}(x_D)]^\top \in \mathbb{R}^{D \times P}.$$

  We have $\boldsymbol{y} = \boldsymbol{F}\boldsymbol{\omega}$.

- Student fits linear model $\boldsymbol{\theta}^\top \boldsymbol{f}(x)$ by least squares:

$$\hat{\boldsymbol{\theta}} = \underline{\boldsymbol{f}}^\dagger \boldsymbol{y} = \underline{\boldsymbol{f}}^\dagger \underline{\boldsymbol{F}} \boldsymbol{\omega}.$$

  Note: gradient descent on least-squares objective $\|\boldsymbol{y} - \boldsymbol{\theta}^\top \underline{\boldsymbol{f}}\|^2$ converges to this.

- Interested in test loss:

$$\mathcal{L} := \mathbb{E}_{x,\omega,x_{1:D}}[(\boldsymbol{\omega}^\top \boldsymbol{F}(x) - \hat{\boldsymbol{\theta}}^\top \boldsymbol{f}(x))^2].$$

- Of particular importance are, resp., the covariance and Gram matrix of the data:

$$\bar{C} := \frac{1}{D}\underline{\boldsymbol{F}}^\top \underline{\boldsymbol{F}} \in \mathbb{R}^{S\times S}, \qquad \bar{K} := \frac{1}{P}\underline{\boldsymbol{f}\boldsymbol{f}}^\top \in \mathbb{R}^{D\times D},$$

$$\bar{\mathcal{K}} := \frac{1}{S}\underline{\boldsymbol{F}\boldsymbol{F}}^\top \in \mathbb{R}^{D\times D}$$

- Resp. in the under- and over- parameterized regimes, we have:
  Under-: $D \gg P \gg 1$: $\bar{C} \approx \mathbb{E}[\bar{C}] =: C$,
  Over-: $P \gg D \gg 1$: $\bar{K} \approx \mathbb{E}[\bar{K}] =: K$, $\bar{\mathcal{K}} \approx \mathbb{E}[\bar{\mathcal{K}}] =: \mathcal{K}$.

- Using these, they arrive at substantially simpler approx for $\mathcal{L}$.
  ... Details - see paper.

- Turns out: leading-order behavior of $\lim_{D\to\infty} \mathcal{L}(D,P)$, $\lim_{P\to\infty} \mathcal{L}(D,P)$ determined by spectrum of $C, \mathcal{K}$.

- When spectrum exhibits power decay,

$$\lambda_n \asymp n^{-1-\alpha_K},$$

then

$$\lim_{D\to\infty} \mathcal{L}(D,P) \propto P^{-\alpha_K}, \qquad \lim_{P\to\infty} \mathcal{L}(D,P) \propto D^{-\alpha_K}.$$

- Fact: when the kernel $k(x,x') = \frac{1}{S}\sum_{i=1}^{S} F_i(x)F_i(x')$ is smooth, the spectrum of the corresponding integral operator (hence of $C, \mathcal{K}$) exhibits power decay. Specifically,

$$\lambda_n \lesssim n^{-1-t/d}$$

when k is t-times continuously differentiable.

- What have we achieved? Unlike previous hand-wavy argument, have a model where scaling law $\propto D^{-c/d}, P^{-c/d}$ is essentially precise.

That's all I have to say. Let's open up the paper and look at some plots.