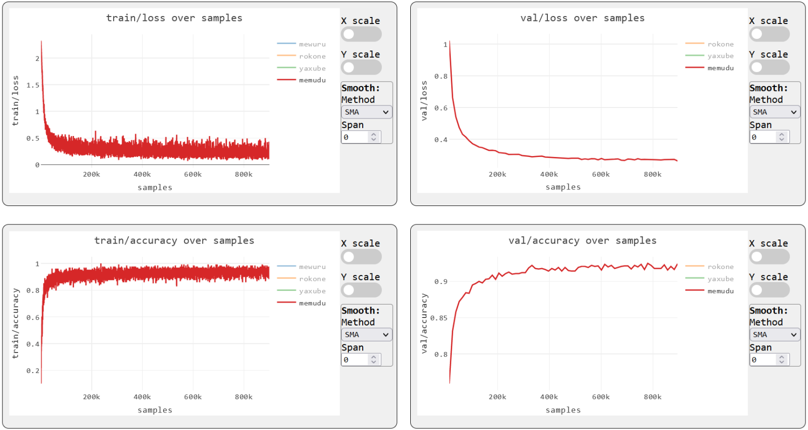
Evaluations and Scaling Laws

Samy Wu Fung, Daniel McKenzie, Michael Ivanitskiy

2024-03-05

Evaluating ML systems



Smooth:
Method SMA
Span 0

Smooth:
Method SMA
Span 0

Smooth:
Method SMA
Span 0

Smooth:
Method SMA
Span 0

Figure 1: How we might evaluate an MNIST-trained DNN

Other evaluation metrics

- histograms for each class
- for boolean identification, ROC (Receiver Operating Characteristic) curve
  - based on some threshold, number of type I vs type II errors
- what else?

Evaluating language models

Perplexity: a measure of how well a language model predicts tokens over a dataset X. Lower perplexity means a model is better at predicting the tokens in X.

$$\texttt{PPL}(X) = \exp\left\{ -\frac{1}{t} \sum_{i}^{t} \log p_\theta \left(x_i \mid x_{<i} \right) \right\}$$

Where $p_\theta(x_i|x_{<i})$ is the probability mass placed on the correct i-th token, given the preceding tokens $x_{<i}$, for a model with parameters θ.

How good are humans at this?

Language Modelling Game

As it turns out, even 125M parameter (GPT-1 sized) models are *better at this than humans*.

See lesswrong.com/posts/htrZrxduciZ5QaCjw/language-models-seem-

Clearly, though, humans are better at tasks than 125M parameter models. So how are we supposed to evaluate them?

The Turing Test

Originally proposed by Alan Turing in 1950, in his paper Computing Machinery and Intelligence, where it was called the “Imitation Game”.

Problems with the Turing test

- highly variable depending on the judges and contestants
- costly and time-consuming to run
- we. . . don't actually want our AI systems to be very good at pretending to be human, do we?
    - most humans are not good at coding, solving differential equations, writing poetry, or doing the sorts of tasks GPT-4 is good at
    - current AI systems are *actively trained to be honest about not being human*, and this is very intentional and does not emerge naturally from pretraining

Beyond the Imiation Game

- Until about ∼2020, the Turing test was mostly hypothetical
    - coincidentally, the annual Loebner Prize for AI chatbots was discontinued in 2020
- after GPT-3 in 2020, it suddenly became a *lot less hypothetical, very fast*
- enter BIG-bench: Beyond the Imitation Game

modern LLM evals

- Figuring out what language models are capable of is now an entire field of research
- hundreds of benchmarks exist, and more are being created all the time
- these benchmarks are often based on tests we give humans: GRE, bar exams, LSAT, IMO problems, programming competitions, etc.
- we *no longer evaluate comparisons with the average human*, and instead compare to humans who are experts in that subject matter

Scaling laws

Key Idea: *As we increase the size of a model and its training data, performance gets better.*

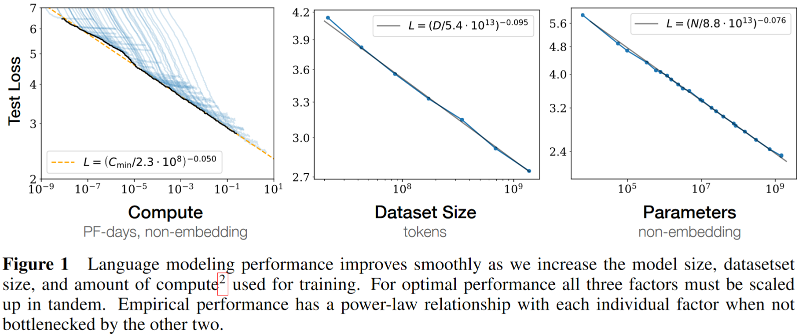


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Figure 2: Scaling Laws for Neural Language Models

For perplexity as defined above, we empirically find that

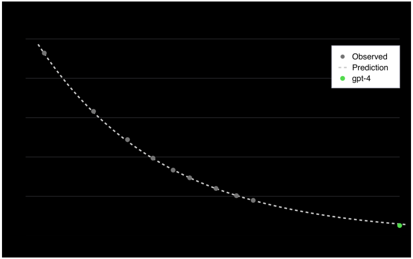
$$L(N, D) = \left(\frac{N_c}{N}\right)^{\alpha_N} + \left(\frac{D_c}{D}\right)^{\alpha_D}$$

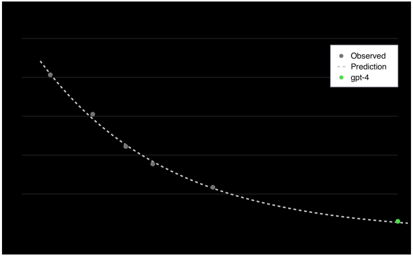
Where:

- L is the perplexity, using cross entropy loss
- N is the number of parameters (excluding embeddings)
- D is the size of the training data (number of tokens)
- $N_c, D_c, \alpha_N, \alpha_D$ are constants

Note that the training compute $C \propto N \cdot D$.

Not only does next token prediction follow predictable curves, but performance on other tasks does as well:





See “Will Scaling Work?” for a more in-depth analysis.

What this means

- if we had infinite compute and infinite data, the trends seem to suggest that these models could be arbitrarily good at predicting tokens
- being good at predicting tokens seems to make you good at pretty much everything we can measure
- finetuning/RLHF makes this even better

compute and data aren't infinite

- in late 2022, it was predicted that we would run out of "high quality" training data by mid-2024
  - "high quality" meaning books, wikipedia articles, academic papers, etc
- low quality training data and non text training data (images, video) we have plenty of, but they seem to be less important for teaching the model about how to be good at science, math, and other academic subjects
- GPUs are a limited and very expensive resource

but…

- even models trained to be *as smart as the average human* would have significant economic and social implications
- *synthetic data* for training LLMs, or data generated/filtered *by other LLMs*, seems to work?
    - this is a developing area, and much of the research is not out in the open
- everyone is racing to build more GPUs (See: Sam Altman asking for $7 *trillion* to build chips)

consequences

We get a new kind of LLM