A Simplistic Model of Neural Scaling Laws: Multiperiodic Santa Fe Processes

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

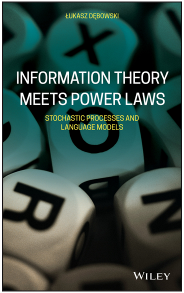
Institute of Computer Science
Polish Academy of Sciences

Math Machine Learning seminar MPI MiS + UCLA
25th May 2023

My interests in statistical language models

- I did my masters in theoretical physics (statistical mechanics).
- Later I worked in statistical natural language processing (part-of-speech tagging of the IPI PAN corpus of Polish).
- But my heart was taken by power laws and information theory (Zipf's law, Hilberg's hypothesis, refutation of finite-state models).

- I did my PhD in information theory and stochastic processes with long memory.
- Ever since then I have been working on mathematical foundations of statistical language modeling (measure theory, ergodic decomposition, excess entropy, Kolmogorov complexity, universal coding and universal prediction).
- **Quite a lot of pretty abstract math...**



ŁUKASZ DĘBOWSKI

INFORMATION THEORY MEETS POWER LAWS

STOCHASTIC PROCESSES AND LANGUAGE MODELS

WILEY

Large language models surprised me, too!

ChatGPT: A self-referential task (optimistic)

Please write a rhymed poem about large language models.

In the realm of words, where knowledge takes flight,
Language models shine, wisdom's beacon bright.
Born from data, their minds vast and grand,
Unleashing power, like an orchestra's command.

From science to art, they delve deep and explore,
Unlocking mysteries, leaving none ignored.
Their language dances, poetic and sublime,
Crafting verses, in rhythm and rhyme.

Yet, remember, they are tools, not the source,
Human minds ignite, creativity's force.
In this dance of machine and art, we find,
The synergy of minds, forever intertwined.

ChatGPT: A self-referential task (pessimistic)

Can you write a verse with a catastrophic tone?

In the realm of words, darkness unfurls,
Language models, a catastrophic whirl.
Born from vast data, their power untamed,
Unleashing chaos, leaving minds maimed.

From science to art, they delve too deep,
Unraveling secrets, in nightmares we sleep.
Their language, a tempest, words fraught with dread,
Crafting verses, where sanity is shred.

Beware their allure, the peril they bear,
Humanity's essence, consumed by their snare.
For in this abyss, where shadows take form,
Our demise awaits, engulfed by the storm.

How Do They Work?

Language models — Cross entropy

Let us write text $(x_1, x_2, \ldots, x_T)$ as x_1^T.

A language model is a (probability) measure on tokens:

$$Q(x_t|x_{t-M}^{t-1}) \geq 0, \quad \sum_{x_t} Q(x_t|x_{t-M}^{t-1}) = 1.$$

The cross entropy of the model is the mean minus log-probability:

$$-\frac{1}{T}\sum_{t=1}^{T} \log Q(x_t|x_{t-M}^{t-1}) \geq 0.$$

It is the average surprisal of model Q on text x_1^T.

We seek for Q that is a computable function of training data x_1^T and minimizes cross entropy on different data, called the test data.

Language models — Embeddings and transformers

In language models based on transformers, probabilities $Q(x_t|x_{t-M}^{t-1})$ are computed by stacking two mechanisms:

- embeddings — vectors x_t corresponding to words/concepts,
- attention — a nonlinear operation on embeddings

$$y_t = \sum_{s=t-M}^{t-1} \frac{\exp(x_t \cdot x_s)}{\sum_{r=t-M}^{t-1} \exp(x_t \cdot x_r)} x_s.$$

The GPT-3 language model:

- Number of parameters: $N = 175$ billions (800 GB RAM).
- Context length: $M = 2048$ words.
- Training data: Common Crawl (410 bln, 60%), WebText2 (19 bln, 22%), books (67 bln, 16%), Wikipedia (3 bln, 3%).

Power Laws in Language and in Language Models

Zipf-Mandelbrot's and Herdan-Heaps' law

Shakespeare's
First Folio/35 Plays:

rank $r(w)$	freq $f(w)$	word w
1	21557	I
2	19059	and
3	16571	to
4	14921	of
5	14491	a
6	12077	my
7	10463	you
8	9789	in
9	8754	is
10	7428	that
...	...	...

Numbers of tokens and types:

$$N = \sum_w f(w), \quad V = \sum_w 1.$$

Zipf-Mandelbrot's law:

$$r(w) \approx \frac{V}{f(w)^\beta}, \quad \beta \in (0,1).$$

Herdan-Heaps' law:

$$V \propto N^\beta, \quad \beta \in (0,1).$$

Power laws in language models

$\boldsymbol{Q}(\boldsymbol{N},\boldsymbol{T})$ — model with $\boldsymbol{N}$ parameters trained on $\boldsymbol{T}$ tokens.

$\mathcal{L}(\boldsymbol{N},\boldsymbol{T})$ — cross entropy of $\boldsymbol{Q}(\boldsymbol{N},\boldsymbol{T})$ on the test data.

Kaplan et al. (2020) observed empirically that

$$\mathcal{L}(N,T) \approx \left[\left(\frac{N_0}{N}\right)^{\frac{\gamma_N}{\gamma_T}} + \frac{T_0}{T}\right]^{\gamma_T} \approx \left(\frac{N_0}{N}\right)^{\gamma_N} \vee \left(\frac{T_0}{T}\right)^{\gamma_T}$$

for $N_0 = 6.4 \times 10^{13}$, $T_0 = 1.8 \times 10^{13}$, $\gamma_N = 0.076$, $\gamma_T = 0.103$.

The more data and the more parameters, the better is the model:

$$\mathcal{L}(\infty,T) \approx \left(\frac{T_0}{T}\right)^{\gamma_T}, \quad \mathcal{L}(N,\infty) \approx \left(\frac{N_0}{N}\right)^{\gamma_N}, \quad \mathcal{L}(\infty,\infty) \approx 0.$$

For each T there is roughly an optimal $N = N_0(T/T_0)^{\gamma_T/\gamma_N}$.

A Toy Language Model

The goal

We will exhibit a toy model of data such that

$$\mathbb{E}\,\mathcal{L}(N,T) \approx \left(\frac{N_0}{N}\right)^{\gamma_N} \vee \left(\frac{T_0}{T}\right)^{\gamma_T},$$

where for an arbitrary $c > 0$, we have underparameterization

$$\gamma_N = \frac{1}{c} > \gamma_T = \frac{1}{c+1}.$$

Note that Kaplan et al. (2020) observed overparameterization

$$\gamma_N < \gamma_T.$$

The optimal number of parameters is $N = N_0(T/T_0)^{\gamma_T/\gamma_N}$.

A toy model of language — Santa Fe processes (2002)

Santa Fe processes are sequences $(X_t)_{t\in\mathbb{N}}$ of pairs

$$X_t = (K_t, Z_{K_t})$$

where $(K_t)_{t\in\mathbb{N}}$, called narration, is a sequence of natural numbers and $(Z_k)_{k\in\mathbb{N}}$, called knowledge, is a sequence of coin flips.

A semantic interpretation

Process $(X_t)_{t\in\mathbb{N}}$ is a sequence of propositions describing knowledge $(Z_k)_{k\in\mathbb{N}}$ at random but consistently:

- Proposition $X_t = (k, z)$ asserts that the k-th coin flip is z, in such way that one can determine both k and z.
- For $X_t = (k, z)$ and $X_s = (k', z')$ we do not know in advance which coin flips they describe but $k = k' \implies z = z'$.

$\implies$ An information-theoretic explanation of Zipf's law! (2011)

Narration model — Multiperiodic sequences (2023)

A multiperiodic sequence:

$$1, 2, 1, 3, 1, 4, 1, 2, 1, 5, 1, 6, 1, 2, 1, 3, 1, 7, 1, 2, 1, 8, 1, 4, 1, 2, 1, \ldots$$

The rule of generation:

If we delete tokens < 1, type 1 appers every $\pi_1 = 2$ tokens.
If we delete tokens < 2, type 2 appers every $\pi_2 = 3$ tokens.
If we delete tokens < 3, type 3 appers every $\pi_3 = 4$ tokens.
...
If we delete tokens $< r$, type r appers every $\pi_r = r + 1$ tokens.

Multiperiodic sequences — The algorithm

Infinite Prague Jewish Clock:

```
Require: List $\pi[r] \in \mathbb{N}$ for $r \in \mathbb{N}$.            ▷ periods
Require: List $\phi[r] = 1$ for $r \in \mathbb{N}$.                      ▷ hands
Ensure: List $k[t] \in \mathbb{N}$ for $t \in \mathbb{N}$.               ▷ multiperiodic sequence
for $t \in \mathbb{N}$ do
    $r_{\text{active}} := 0$
    $r_{\text{iter}} := 1$
    while $r_{\text{active}} = 0$ do
        if $\phi[r_{\text{iter}}] > 1$ then
            $\phi[r_{\text{iter}}] := \phi[r_{\text{iter}}] - 1$
        else
            $r_{\text{active}} := r_{\text{iter}}$
        $r_{\text{iter}} := r_{\text{iter}} + 1$
    $\phi[r_{\text{active}}] := \pi[r_{\text{active}}]$
    $k[t] := r_{\text{active}}$
```

If we initialize $\phi_r \sim \text{Unif}(1, 2, ..., \pi_r)$, we obtain a stationary ergodic process with a zero entropy rate for $\pi_r \leq cr$ with $c > 0$.

Multiperiodic sequences — Relative frequency

The relative frequency of types $\geq r$:

$$f_r := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} 1\{k_t \geq r\}$$
$$= \left(1 - \frac{1}{\pi_1}\right)\left(1 - \frac{1}{\pi_2}\right) \dots \left(1 - \frac{1}{\pi_{r-1}}\right)$$

Example

Let $\pi_r \approx cr$ for some $c > 0$ and all $r \in \mathbb{N}$. We may estimate

$$f_r \approx \exp \sum_{i=1}^{r-1} \log\left(1 - \frac{1}{ci}\right) \approx \exp \int_1^r \log\left(1 - \frac{1}{cx}\right) dx$$
$$\approx \exp\left(-\int_1^r \frac{dx}{cx}\right) = \exp\left(-\frac{\log r}{c}\right) = r^{-1/c}.$$

The waiting time and the number of types

The waiting time and the number of types:

$$w_r := \min\{t \in \mathbb{N} : k_t = r\} \geq r$$
$$n_t := \#\{k_1, k_2, \dots, k_t\} = \max\{r \in \mathbb{N} : w_r \leq t\} \leq t$$

A sandwich bound that resembles the Kac lemma:

$$\frac{1}{f_r} \leq w_r < \sum_{j=1}^{r} \frac{1}{f_j}$$

Example

Let $\pi_r \approx cr$ for some $c > 0$ and all $r \in \mathbb{N}$. We have

$$w_r \sim r^{(c+1)/c}, \qquad n_t \sim t^{c/(c+1)}.$$

A Toy Model of Learning

Multiperiodic Santa Fe process — Model of learning

Environment:
A learning agent observes $(X_t)_{t\in\mathbb{N}}$ with $X_t=(k_t,Z_{k_t})$, where narration $(k_t)_{t\in\mathbb{N}}$ is a known multiperiodic sequence and knowledge $(Z_k)_{k\in\mathbb{N}}$ is a sequence of independent coin flips.

Goal:
The learning agent has to read first T data points X_1^T, then to compute N binary parameters $B_1^N=g_1(X_1^T;N)$, and finally to predict the remaining sequence as $\hat{X}_{T+i}=g_2(T+i;B_1^N)$.

Risk: We want to minimize the error rate

$$\mathcal{L}(N,T):=\lim_{I\to\infty}\frac{1}{I}\sum_{i=1}^{I}1\left\{X_{T+i}\neq\hat{X}_{T+i}\right\}.$$

Multiperiodic Santa Fe process — Reasonable learner

Reasonable parameters:

Parameters B_1^N should be chosen as the optimal estimators of coin flips Z_1^N. If token (r, Z_r) appears in data X_1^T, setting $B_r = Z_r$ can be actually carried out. If token (r, Z_r) does not appear in data X_1^T then we may put $B_r = 0$. In this way, we obtain

$$B_r = \begin{cases} Z_r, & r \le N \wedge n_T, \\ 0, & r > N \wedge n_T. \end{cases}$$

We apply notation $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$.

Reasonable predictors:

Some reasonable predictors are

$$\hat{X}_{T+i} = (k_{T+i}, B_{k_{T+i}}).$$

Multiperiodic Santa Fe process — Error rate

Hence, the error rate is the relative frequency of $(Z_{k_{T+i}} \neq B_{k_{T+i}})$,

$$\mathcal{L}(N,T) = \lim_{I\to\infty} I^{-1}\sum_{i=1}^{I} 1\{Z_{k_{T+i}} \neq B_{k_{T+i}}\}.$$

Averaging over random knowledge $(Z_k)_{k\in\mathbb{N}}$, we derive

$$\begin{aligned}\mathbb{E}\,\mathcal{L}(N,T) &= \lim_{I\to\infty} I^{-1}\sum_{i=1}^{I} P(Z_{k_{T+i}} \neq B_{k_{T+i}}) \\ &= \frac{1}{2}\lim_{I\to\infty} I^{-1}\sum_{i=1}^{I} 1\{k_{T+i} > N\wedge n_T\} = \frac{f_{N\wedge n_T}}{2}.\end{aligned}$$

Example

Let $\pi_r \approx cr$. We have $f_r \sim r^{-1/c}$ and $n_t \sim t^{c/(c+1)}$. Hence

$$\mathbb{E}\,\mathcal{L}(N,T) \approx \left[\frac{N}{N_0}\wedge\left(\frac{T}{T_0}\right)^{\frac{c}{c+1}}\right]^{-\frac{1}{c}} = \left(\frac{N_0}{N}\right)^{\frac{1}{c}} \vee \left(\frac{T_0}{T}\right)^{\frac{1}{c+1}}.$$

Is This Relevant?

Something seems quite relevant but what?

1. Santa Fe processes were independently reinvented by:
    - M. Hutter. Learning curve theory. `https://arxiv.org/abs/2102.04074`, 2021.
    - E. J. Michaud, Z. Liu, U. Girit, M. Tegmark. The Quantization Model of Neural Scaling. `https://arxiv.org/abs/2303.13506`, 2023
2. Hutter wrote about this sort of simplistic models:

   The toy model studied in this work is admittedly totally unrealistic as a Deep Learning model, but we believe it captures the (or at least a) true reason for the observed scaling laws w.r.t. data.

   Unfortunately, he did not develop a discussion of this intuition.

Santa Fe decomposition

- When we read a text in natural language, we may feel that it consists of contiguous propositions describing discrete facts.
- Since there are only countably many distinct propositions $\boldsymbol{x_t}$ and countably many distinct mentioned facts $\boldsymbol{b_k}$, we may enumerate them by natural numbers and arrive at a representation of individual propositions $\boldsymbol{x_t} = (\boldsymbol{k_t}, \boldsymbol{b_t})$ that resembles Santa Fe decomposition $\boldsymbol{x_t} = (\boldsymbol{k_t}, \boldsymbol{z_{k_t}})$.
- Two delicate questions are:

— Can decompositions $(\boldsymbol{k_t}, \boldsymbol{b_t})$ be effectively computed?

— Does $\boldsymbol{k_t} = \boldsymbol{k_{t'}}$ imply $\boldsymbol{b_t} = \boldsymbol{b_{t'}}$?

Only then we may define immutable facts $\boldsymbol{z_r} := \boldsymbol{b_t}$ for $\boldsymbol{k_t} = \boldsymbol{r}$.

- But even if $\boldsymbol{k_t} = \boldsymbol{k_{t'}}$ implies $\boldsymbol{b_t} = \boldsymbol{b_{t'}}$ only for time indices $\boldsymbol{t}$ and $\boldsymbol{t'}$ that are close enough then the text still exhibits some properties of the Santa Fe process.

Conditional determinism of narration

- The Santa Fe decomposition posits that text $(x_t)_{t\in\mathbb{N}}$ is a composition of knowledge $(z_k)_{k\in\mathbb{N}}$ and narration $(k_t)_{t\in\mathbb{N}}$.
- Is there a good reason to suspect that the narration is deterministic given the knowledge and resembles the multiperiodic process?
- Determinism of narration is equivalent to zero entropy rate and, as widely known, Shannon (1951) showed that the entropy rate of natural language is 1 bit per letter.
- There have been researchers like Hilberg (1990), looking at the same data and claiming the zero entropy rate.
  — cube-logarithmic growth of the maximal repetition (2015)
- The stake is high and it is better to stay cautious.

Tampering with the infinite clock mechanism

- The Infinite Prague Jewish Clock seems an interesting model for combining determinism and randomness in narration.
- We may tamper with hands ϕ_r: set them at random values, reset them with certain probabilities, introduce correlations.
- All of this can make the output sequence $(k_t)_{t\in\mathbb{N}}$ more similar to the rhythm of daily chores or human utterances:
  - there may be cycles of varying time scales,
  - there may be repetitions,
  - there may be hierarchical structures,
  - there may be bursts and lulls,
  - there may be some residual randomness.
- The open problem seems to learn the true dynamics of hands ϕ_r. Is it a more transparent approach to artificial intelligence than transformers?

Conclusion

Conclusion

- We have a model that reproduces the neural scaling law.
- The model applies Santa Fe processes, multiperiodic sequences, and memory-based learning.
- Probably, it is the simplest model with the vanishing entropy rate and power-law learning curves.
- The neural scaling law seems linked to quantitative linguistic laws such as Zipf-Mandelbrot's and Herdan-Heaps' laws.
- Our toy model predicts underparameterization $\gamma_T < \gamma_N$.
- The serious challenge is to explain why overparameterization $\gamma_T > \gamma_N$, discovered to be beneficial in machine learning, does not impede generalization.
- Sole information theory seems a too weak tool to analyze it.

Further reading

M. Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. `https://arxiv.org/abs/2105.14368`, 2021.

Ł. Dębowski. *Information Theory Meets Power Laws: Stochastic Processes and Language Models*. Wiley & Sons, 2021.

Ł. Dębowski. A simplistic model of neural scaling laws: Multiperiodic Santa Fe processes. `https://arxiv.org/abs/2302.09049`, 2023.

M. Hutter. Learning curve theory. `https://arxiv.org/abs/2102.04074`, 2021.

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. `https://arxiv.org/abs/2001.08361`, 2020.

A. Maloney, D. A. Roberts, and J. Sully. A solvable model of neural scaling laws. `https://arxiv.org/abs/2210.16859`, 2022.

E. J. Michaud, Z. Liu, U. Girit, and M. Tegmark. The quantization model of neural scaling. `https://arxiv.org/abs/2303.13506`, 2023.

D. A. Roberts, S. Yaida, and B. Hanin. *The Principles of Deep Learning Theory*. Cambridge University Press, 2022.