Neural Networks Learn Statistics of Increasing Complexity

(Belrose, Fern, Pope, ...)

Distribution simplicity bias → Models learn 1°, 2nd order moments of data first

→ NN perform well on max-entropy distrib whose low order statistics match does of the training set AT FIRST

[it's possible to poison a NN to achieve "ϕ" loss but still behave randomly on a test set]

→ distributional simplicity bias

What we know already

→ Refinetti (2023) → sequence of synthetic datasets with increasingly approximation to the real data, first check points work well too.

Methods

→ Train on real dataset and the use synthetic data to probe reliance on statistics of different orders.

Theory & Method

$\mathcal{L}(x)$ the loss of a NN on input x.
If $\mathcal{L}(x)$ is ANALYTIC, we can Taylor expand the loss $\forall x$ around μ:

$$\mathcal{L}(x) = \sum_{\alpha \in \mathbb{N}^d} \frac{(x-\mu)^\alpha}{\alpha!} (\partial^\alpha \mathcal{L})(\mu)$$

→ MULTI-INDEX

if $\alpha = (1, 4, 6) \rightarrow (X-\mu)^\alpha = (X_1-\mu_1)(X_2-\mu_2)^4 (X_3-\mu_3)^6$

if $\alpha! = 1!\,4!\,6!$

If x comes from a distrib of COMPACT support (true for text), we take:

$$E[\mathcal{L}(x)] = \sum_{\alpha \in \mathbb{N}^d} \frac{\partial^\alpha \mathcal{L}}{\alpha!}(\mu)\, E[(x-\mu)^\alpha]$$

→ expected loss

→ moments of data distribution

What do we expect? → "Grafting" low order stat of class B onto class A should cause (1) the model to treat examples of A as B.

→ "Deleting" the info of high order statistics shouldn't be harmful

Optimal Transport for (1)

→ Transforming samples from

PROB DISTR 1 → PROB DISTR 2

minimizing avg distance that samples are moved

OT methods → coordinate-wise quantile normalization (1° order)
→ bounded shift (1° order)
→ Gaussian OT (1°/2° order)

Gaussian OT

Given $P = \mathcal{N}(\mu_P, \Sigma_P)$, $Q = \mathcal{N}(\mu_Q, \Sigma_Q)$, the map $T(x) = A(x - m_P) + m_Q$ is the OT map from P to Q, with:

$$(3)\quad A = \Sigma_P^{-1/2} \left(\Sigma_P^{1/2} \Sigma_Q \Sigma_P^{1/2}\right)^{1/2} \Sigma_P^{-1/2}$$

(Given K image classes, each $C \times H \times W$, compute μ & cov, plug in (3) and)

Coordinate wise Quantile Normalization (CQN)

To make 2 SCALAR RANDOM VARIABLES identical in their statistical properties

How it works? → X r.v has $F_X(x)$. Then $F_X(\tilde{X})$ will have $\sim U(0,1)$

Maximum entropy sampling for (2)

To construct pdf using partial knowledge and lowering knowledge in HO stats → the low order statistics derived from a TRAINING DATASET

We want to construct the Max entropy distr P and sample from it

Which distrib? In $\mathbb{R}^d$, Gaussian (μ, Σ) is MED → (Compact support??)

Discrete domain

"The cat is brown" → [1 8 4 ... 52] → [ ] 1 763

Theorem
n-gram statistics are moments

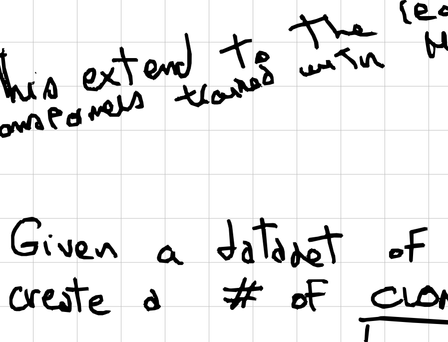
Theorem
Equal embedding moments

Language modeling

14M → 10 seed
70M → 10 seed
160M → 10 seed
410M → 5 seed

- n-gram language models: compute token unigram & bigram frequencies across Pythia Training set and construct ME n-gram LM.

Results → unigram sequence loss reaches its lowest point before bigram sequence loss and higher min value



→ UNIGRAMS ARE LEARNED FIRST & THEN THE BIGRAMS

"A distributional simplicity bias in Learning Dynamics of Transformers" (Laio, Gerace, ...)

NN trained on SGD learn using simplicity biases
↓
DNN learn INCREASINGLY COMPLEX APPROXIMATIONS OF THE DISTRIB of DATA

→ Does this extend to the realm of Transformers trained with MLM?

Main idea: Given a dataset of NL, they create a # of CLONES

WHAT IS
→ each clone approximates the underlying data distribution of WikiText 103 but only including interactions between tokens up to a specific order.

HOW?
→ we use MLM to train a new type of Transformers with MULTIPLE LAYERS OF FACTORED ATTENTION WITH QUADRATIC ACTIVATION FUNCTIONS,
→ The depth of the models controls the degree of interactions
↓
Then sample the model with Monte-Carlo-sampling Techniques

[Also: Cagnetta et al. 28, 29
Garnier-Brun et al.]

Expressing many-body distrib. with Factored attention

$S = (s_1, \ldots, s_L)$ "Lucia è simpatica" (s_0 s_1 s_2)

$s_i \in \{1 \ldots |V|\}$

the sequence $(s_1, \ldots, s_i = \emptyset, \ldots, s_L)$ is given as input to the transformer $\rightarrow (x_1, \ldots, x_L)$

$$(x_1, \ldots x_L) \xrightarrow[\text{EMBEDDING}]{\text{S-ATTENT}} (y_1, \ldots y_L)$$ (CONT. EMB in d-dim) where $y_i \in \mathbb{R}^d$

The transformer is trained to minimize

$$\mathcal{L} = -\sum_{m=1}^{M} \sum_{i\alpha} s_{i\alpha}^{(m)} \log\left(p_{mlm}(s_{i\alpha}^m | s_{\neq i}^m)\right)$$

Where M denotes the total number of examples in the dataset, m indexes indiv. samples

(Rende et al.) considered a simplified attention mechanism FACTORED ATTENTION where ATTENTION WEIGHTS are INPUT INDEPENDENT

$$y_{i\alpha} = \sum_{j=1}^{L} A_{ij} \sum_{\beta=1}^{|V|} V_{\alpha\beta}\, x_{j\beta}$$

$A \in \mathbb{R}^{L \times L}$ & $V \in \mathbb{R}^{d \times d}$ are matrices of Trainable params

This way, the PREDICTION OF A SINGLE-HEAD ATTENTION FOR THE MASKED TOKEN

$$p_{mlm}(s_{i\alpha} = 1 | s_{\neq i}) = \text{softmax}\left(\sum_{j \neq 1}^{L} A_{ij} \sum_{\beta=1}^{|V|} V_{\alpha\beta}\, s_{j\beta}\right)$$

It can be shown analitically that, when training a single layer of factored attention via MLM, THE MODEL IS CAPABLE of exactly inferring 2-BODY interactions among INPUT TOKENS, when sampled according to:

$$p(s) = \frac{1}{Z} \exp\left(-\sum_{ij}\sum_{\alpha\beta} J_{ij} \left(U_{\alpha\beta}\, s_{i\alpha}\, s_{j\beta}\right)\right)$$

→ normalizes

→ interaction between position

→ similarities between tokens

SEQUENTIAL LEARNING in NLP

Dataset: TinyStories
Model: BERT - GPT2

→ We do not know the underlying distrib over words, we train the model (*) with varying # of layers on TinyStories → can be sampled to get TinyStories

Clones → we exploit the generative models in (*) to generate the clones of the TinyStories data set

$F(x) = x^2$

→ the clones are meant to approximate the TinyStorie distribution with increased Fidelity

We use $n = 2, 4, 6$ layers of factored attention to generate clones that include effective interactions up to 3, 9 and 33

$|V| = 10000$
↓
we perform a Finite-rank approximation of the value matrix

The Final value of the test loss decreases systematically with the order of interaction modelled by architecture

Sampling Procedure

Sampling process of Goyal et al. based on Metropolis-Hastings algorithm.

→ First select 620 examples of the TEST SET at random

$$(s_1, \ldots, s_L) \rightarrow E(s_1, \ldots s_L) = -\sum_{i=1}^{L} h_i \cdot s_i$$

Define its score

LOGITS

Then a MH Sampler is run on the joint distrib

$$p(s_1, \ldots s_L) \propto \exp[-E(s_1, \ldots s_L)]$$

$t_0 : (s_1^{t=\emptyset}, \ldots, s_L^{t=\emptyset})$

t_i : – i-th Token is masked & replace $s_i^{t+1} \sim p_{mlm}(s_i | s_{\neq i}^t)$
– MH criterion of acceptance

Validating the clones

Checks with a standard 4-layer Transformer encoder on T-S

AAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAA
AAAAAAA