A Probabilistic Interpretation of Transformers

Alexander Shim
ML Collective
alex.shim@gmail.com

Abstract:
We propose a probabilistic interpretation of exponential dot product attention of transformers and contrastive learning based off of exponential families. The attention sublayer of transformers is equivalent to a gradient ascent step of the log normalizer, which is the log-sum-exp term in the Hopfield theory of attention. This ascent step induces a parallel expansion of points, which is counterbalanced by a contraction from layer normalization. We also state theoretical limitations of our theory and the Hopfield theory and suggest directions for resolution.

Transformer Attention

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Exponential Families:

$$p(x|\eta) = \frac{1}{Z(\eta)}h(x)e^{u(x)\cdot\eta}$$
$$\propto e^{u(x)\cdot\eta}$$
$$Z(\eta) := \int h(x)e^{u(x)\cdot\eta}dx$$

$u(x)$ is the sufficient statistic
η is the natural parameter
$h(x)$ is the intrinsic measure
$Z(\eta)$ is the normalizer

Exponential family property

$$\nabla_\eta \log Z(\eta) = E_{x\sim p(x|\eta)}\left[u(x)\right]$$

Exponential dot product attention

$$p(x = key_i|\eta = query) = \frac{e^{key_i\cdot query}}{\sum_j e^{key_j\cdot query}}h(x)$$

$$h(x) = \begin{cases} 1 & x = key_j \text{ for some j} \\ 0 & \text{otherwise} \end{cases}$$

Discrete intrinsic measure — Attention sublayer

$$\nabla_\eta \log Z(\eta) = \sum_i P(x_i|\eta)x_i \quad \sum_i A(key_i, query)\, value_i$$
$$= \sum_i A(x_i,\eta)x_i \quad \sum_i \frac{e^{key_i\cdot query}}{\sum_j e^{key_i\cdot query}} value_i$$

Attention performs a gradient update on the log normalizer

Gradient updates for different distributions

Proposition 1. *Let $X = R^D$, $h : X \to R^+$.*

(a) If $h(x) := \sum_{n=1}^N \delta(x = x_n)$, then $\nabla_\eta \log \int h(x)e^{x^T\eta}dx = \sum_{n=1}^N \frac{e^{x_n^T\eta}}{\sum_{n'} e^{x_{n'}^T\eta}} x_n$.

(b) If $h(x) = p_0(x|\eta_1,\eta_2)$, where $p_0(x|\eta_1,\eta_2)$ is the exponential family distribution $p_0(x|\eta_1,\eta_2) = \frac{1}{Z_0(\eta)})h_0(x)e^{x^T\eta}e^{u_2(x)^T\eta_2}$, with sufficient statistic $u_1(x) = x$ and arbitrary sufficient statistic $u_2(x)$, natural parameters (η_1,η_2), intrinsic measure $h_0(x)$, and normalizer $Z_0(\eta_1,\eta_2) = \int h_0(x)e^{x^T\eta_1}e^{u_2(x)^T\eta_2}dx$, then $\nabla_\eta \log \int h(x)e^{x^T\eta} = E_{x\sim p_0(x|\eta_1+\eta,\eta_2)}[x]$

Corollary 1. *If $h(x) = \mathcal{N}(x;\mu,\Sigma)$, then $\nabla_\eta \log \int h(x)e^{x^T\eta}dx = \mu + \Sigma\eta$.*

Proposition 2. *If $h(x) = \sum_{n=1}^N \pi_n \mathcal{N}(\mu_n,\Sigma)$, where $\pi_n \in R^+$, then $\nabla_\eta \log \int h(x)e^{x^T\eta}dx = \Sigma\eta + \sum_{n=1}^N \frac{\pi_n e^{\mu_n^T\eta}}{\sum_{n'=1}^N \pi_{n'}e^{\mu_{n'}^T\eta}}\mu_n$*

Why Optimize log Z?

$$\log Z(\eta) := \log \int h(x)e^{u(x)\cdot\eta}dx \text{ (continuous)}$$
$$\log Z(\eta) := \log\left(\sum_{i=1}^N e^{u(x_i)\cdot\eta}\right) \text{ (discrete)}$$

In convex optimization (Boyd, 2009), log-sum-exp can be solved by Lagrange multipliers

$$\begin{array}{ll} \text{minimize} & f_0(y) = \log\left(\sum_{i=1}^m \exp y_i\right) \\ \text{subject to} & Ax + b = y, \end{array}$$

… converting it into a maximum entropy problem, of a multinoulli distribution on a probability simplex

$$\begin{array}{ll} \text{maximize} & b^T\nu - \sum_{i=1}^m \nu_i \log \nu_i \\ \text{subject to} & \mathbf{1}^T\nu = 1 \\ & A^T\nu = 0 \\ & \nu \succeq 0, \end{array}$$

leads to KL divergence, statistical mechanics, Jaynes' max entropy

Information Geometry

$$G(\eta) := \log Z(\eta)$$
$$B_G(\eta_1,\eta_2) = KL\left(p(x|\eta_1)||p(x|\eta_2)\right)$$
$$B_{G^*}(\bar{u}_1,\bar{u}_2) = KL\left(p(x|\eta_2)||p(x|\eta_1)\right)$$

We can view our hidden states as a distribution of distributions, over the natural parameters or the expected sufficient statistics

$$e^{q_i^T k_j} = e^{q_i^T \nabla_\eta \log Z(q_j)}$$

Affine approximation of the link function is given by the Fisher information matrix (covariance)

$$\nabla^2 \log Z(\eta) = \Sigma$$

Proof of Gaussian equilibrium hidden state distribution

$\eta + \nabla_\eta \log Z(\eta)$ defines our pointwise attention update operator
RN is our renormalization operator, such as layer or batch normalization

Theorem 1. *(Informal) $p_{eq}(\eta) = \mathcal{N}(\Sigma^{-1}\mu,\Sigma^{-1})$ is an equilibrium distribution of the renormalization operator RN with mean $\Sigma^{-1}\mu$ and covariance Σ^{-1}, composed with the attention update operator, assuming intrinsic measure $h(x) : \eta \to \Sigma\eta$.*

Comparison to Hopfield Networks is All You Need

- Similar energy function / gradient update, ours has an information theoretic interpretation
- Hopfield patterns / keys are fixed, ours change with the hidden state
- Hopfield ignores weights, ours speculates an information theoretic interpretation
- Neither deals with FC layer or multihead attention well

Conclusion/Future Work

Provides a strong probabilistic foundation, upon which future work, particularly using statistical sampling methods on hybrid distributions and SMC, can build.
Possible observational experiments on how the distribution of hidden states changes over layers and training
Test if query and key matrices of different attention heads form local gaussian approximations, in the spirit of the local linear embeddings roots of contrastive learning.