[

A Probabilistic Interpretation of Transformers

Alexander Shim [1]

]

References

A. Proofs

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

(a) If $h(x) := \sum_{n=1}^{N} \delta(x = x_n)$, then $\nabla_\eta \log \int h(x) e^{x^\intercal \eta} dx = \sum_{n=1}^{N} \frac{e^{x_n^\intercal \eta}}{\sum_{n'} e^{x_{n'}^\intercal \eta}} x_n$.

(b) If $h(x) = p_0(x|\eta_1, \eta_2)$, where $p_0(x|\eta_1, \eta_2)$ is the exponential family distribution $p_0(x|\eta_1, \eta_2) = \frac{1}{Z_0(\eta)} h_0(x) e^{x^\intercal \eta} e^{u_2(x)^\intercal \eta_2}$, 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 $Z_0(\eta_1, \eta_2) = \int h_0(x) e^{x^\intercal \eta_1} e^{u_2(x)^\intercal \eta_2} dx$, then $\nabla_\eta \log \int h(x) e^{x^\intercal \eta} = E_{x \sim p_0(x|\eta_1+\eta, \eta_2)} [x]$

Proof. (a) $Pr\{x = x_n\} = \frac{e^{x_n^\intercal \eta}}{\sum_{n'} e^{x_{n'}^\intercal \eta}}$, so $\nabla_\eta \log Z(\eta) = E_{x \sim p(x|\eta)} [x] = \sum_n \frac{e^{x_n^\intercal \eta}}{\sum_{n'} e^{x_{n'}^\intercal \eta}} x_n$.

(b) We collect exponential terms into a distribution of the same exponential family as $p_0(x|\eta_1, \eta_2)$

$$\begin{aligned} \int p_0(x) e^{x^\intercal \eta} dx &= \frac{1}{Z_0(\eta_1, \eta_2)} \int h_0(x) e^{x^\intercal (\eta_1 + \eta_2)} e^{u_2(x)^\intercal \eta_2} dx \\ &= \frac{Z_0(\eta_1 + \eta, \eta_2)}{Z_0(\eta_1, \eta_2)} \end{aligned} \tag{1}$$

When evaluating the score, the denominator term has no dependence on η

$$\begin{aligned} \nabla_\eta \log \int p(x|\eta_1, \eta_2) e^{x^\intercal \eta)} dx &= \nabla_\eta \log \frac{Z_0(\eta_1 + \eta), \eta_2}{Z_0(\eta_1, \eta_2)} \\ &= \nabla_\eta \log Z_0(\eta_1 + \eta, \eta_2) \\ &= E_{x \sim p(x|\eta_1+\eta, \eta_2)} [x] \end{aligned} \tag{2}$$

□

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

[1]ML Collective. Correspondence to: Alexander Shim <alex.shim@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Proof. The natural parameter form of a Gaussian is $\frac{1}{Z(\eta_1,\eta_2)}e^{\left\langle x,\Sigma^{-1}\mu\right\rangle+\left\langle xx^\intercal,-\frac{\Sigma^{-1}}{2}\right\rangle}$, where $\langle\,.\,,\,.\,\rangle$ denotes the vectorized dot product for both vectors and matrices, with $\eta_1=\Sigma^{-1}\mu$. Conversely, $\mu=\Sigma\eta_1$.
Hence, $E_{x\sim p(x|\Sigma^{-1}\mu+\eta,-\frac{\Sigma^{-1}}{2})}[x]=\mu+\Sigma\eta$ □

Proposition 2. *If* $h(x)=\sum_{n=1}^N\pi_n\mathcal{N}(\mu_n,\Sigma)$, *where* $\pi_n\in R^+$, *then* $\nabla_\eta\log\int h(x)e^{x^\intercal\eta}dx=\Sigma\eta+\sum_{n=1}^N\frac{\pi_n e^{\mu_n^\intercal\eta}}{\sum_{n'=1}^N\pi_{n'}e^{\mu_{n'}^\intercal\eta}}\mu_n$

Proof. For a Gaussian $\mathcal{N}(x;\mu,\Sigma)$ for $x\in R^d$, the exponential family normalizer is $(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}e^{\frac{1}{2}\mu^\intercal\Sigma^{-1}\mu}$.

$$\begin{aligned}\int\sum_{n=1}^N\pi_n\mathcal{N}(x;\mu_n,\Sigma)e^{x^\intercal\eta}dx&=\sum_{n=1}^N\frac{1}{Z(\mu_n,\Sigma)}\pi_n\int e^{\left\langle xx^\intercal,-\frac{1}{2}\Sigma^{-1}\right\rangle+\left\langle x,\Sigma^{-1}\mu_n\right\rangle+x^\intercal\eta}dx\\&=\sum_{n=1}^N\pi_n\frac{Z(\mu_n+\Sigma\eta,\Sigma)}{Z(\mu_n,\Sigma)}\\&=\sum_{n=1}^N\pi_n e^{\frac{1}{2}(\mu_n+\Sigma\eta)^\intercal\Sigma^{-1}(\mu_n+\Sigma\eta)-\frac{1}{2}\mu_n^\intercal\Sigma^{-1}\mu_n}\\&=e^{\frac{1}{2}\eta^\intercal\Sigma\eta}\sum_{n=1}^N\pi_n e^{\mu_n^\intercal\eta}\end{aligned}\tag{3}$$

The gradient of the log normalizer simplifies to

$$\nabla_\eta\log Z(\eta)=\Sigma\eta+\nabla_\eta\log\sum_{n=1}^N\frac{\pi_n e^{\mu_n^\intercal\eta}}{\sum_{n'=1}^N\pi_{n'}e^{\mu_{n'}^\intercal\eta}}\mu_n\tag{4}$$

which when π_n are uniform $\forall n$ is the sum of the discrete attention update and the $\Sigma\eta$ term. □

If our intrinsic measure is a mixture of Gaussians with different covariance matrices, then we can perform a gradient update on a lower bound, $G_{LB}(\eta)$, of the log normalizer.

Proposition 3. *If* $h(x)=\sum_{n=1}^N\pi_n\mathcal{N}(\mu_n,\Sigma_n)$, *then there exists a lower bound* $G_{LB}=\sum_{n=1}^N\pi_n\left(\frac{1}{2}\eta^\intercal\Sigma\eta+\mu_n^\intercal\eta\right)\le\log Z(\eta)$, *with gradient update* $\nabla_\eta G_{LB}(\eta)=\sum_{n=1}^N\pi_n(\mu_n+\Sigma_n\eta)$

Proof. Using Jensen's inequality on the concave logarithm function,

$$\log\int\sum_{n=1}^N\pi_n\mathcal{N}(x;\mu_n,\Sigma_n)\,e^{x^\intercal\eta}dx\ge\sum_{n=1}^N\pi_n\log\int\mathcal{N}(x;\mu_n,\Sigma_n)e^{x^\intercal\eta}dx\tag{5}$$

From (4),

$$=\sum_{n=1}^N\pi_n\left(\frac{1}{2}\eta^\intercal\Sigma_n\eta+\mu_n^\intercal\eta\right)\tag{6}$$

If we perform a gradient update on the lower bound, we have

$$\nabla_\eta G_{LB}(\eta)=\sum_{n=1}^N\pi_n(\mu_n+\Sigma_n\eta)\tag{7}$$

□

Using Corollary 1, we can apply a generalization of the attention sublayer to update a distribution $p(\eta)$ of natural parameters given a Gaussian intrinsic measure, $h(x) = \mathcal{N}(x; \mu, \Sigma)$. Moreover, if we assume there is a one-to-one transformation of the natural parameters into the intrinsic measure through the activation function, we can prove the stationarity of the attention update operator composed with a renormalization:

Theorem 1. *Suppose the distribution of natural parameters is $p_{eq}(\eta) = \mathcal{N}(\Sigma^{-1}\mu, \Sigma^{-1})$ and we define the intrinsic measure $h(x)$ by a pointwise transformation $x = \Sigma\eta$. Let RN_{η_0,Λ_0} be the renormalization operator such that if $E_{\eta\sim p(\eta)}\left[(\eta, \eta\eta^\intercal)\right] = (\bar{\eta}, \bar{\Sigma})$, RN_{η_0,Λ_0} maps $\eta \to \eta_0 + \Lambda_0^{\frac{1}{2}}\bar{\Sigma}^{-\frac{1}{2}}(\eta - \bar{\eta})$. Then the composition of renormalization operator and the attention update operator $RN_{\Sigma^{-1}\mu,\Sigma^{-1}} \circ A \circ (p_{eq}(\eta), h(x)) = p_{eq}(\eta)$.*

Proof. The attention operator acts pointwise on each η by $\eta' = \eta + \nabla_\eta \log \int h(x) e^{x^\intercal \eta} dx$. We can reparametrize η as $\eta = \Sigma^{-1}\mu + \Sigma^{-\frac{1}{2}}\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. Hence $x = \mu + \Sigma^{\frac{1}{2}}\epsilon$, which is the reparametrization of $\mathcal{N}(\mu, \Sigma)$.
By Corollary 1,

$$\eta' = \mu + (I + \Sigma)\eta \tag{8}$$

This is an affine transformation of the natural parameter, and an affine transformation of a gaussian random variable is a gaussian random variable. The renormalization operator is another affine transformation which remaps it to a distribution with mean and covariance of $p_{eq}(\eta)$, and once again it must be a Gaussian distribution.

□

B. Hopfield Networks is All You Need comparison

One key difference is that for the Hopfield theory, the patterns are fixed, while for our theory the patterns are the changing hidden states. Notably later versions of the Hopfield paper added Hopfield layers with changing patterns, but theoretical proofs of the limiting behavior apply for fixed patterns. Arguably, the Hopfield analysis was not meant to answer how the attention sublayers transformed the input, and the focus on exponential storage capacity focused more on the ability of attention to quickly converge from noisy inputs to stored patterns.

Notably, while the encoder and decoder involve attention sublayers which use the hidden states as patterns, fitting our interpretation better, the decoder layers also add a second attention block which apply attention to the encoder input, fitting the Hopfield paper conditions.

From a conceptual perspective, the advantage of the original Hopfield network was that the patterns were stored in the weights. For modern Hopfield networks, the patterns must be stored separately, and they are not directly related to the transformer weights, removing the benefit of biological plausibility.

The Hopfield theory requires a more specific set of assumptions, which justify the energy function, whereas our theory is based off of information theoretic ideas of a widely used free energy objective, which is dual to maximum entropy. We argue that our theory is more directly responsible for the theoretic interpretation of the Hopfield energy function and gradient updates and it does not rely on motivating factors behind Hopfield networks.

One mutual deficiency for both theories is that they do not explain the transformer attention weight transformations nor does it explain multihead attention. Not having to consider weight transformations, multihead attention, or FC layers makes equilibrium analysis much simpler. Unfortunately, the only parameters in the attention sublayer are in these weight transformations, requiring all knowledge of the data distribution to be encoded into these parameters - otherwise we might expect our language models to perform fairly well without any weights. However, it may be possible that some weights may be less sensitive or at least easier to train than others - linearly transforming the key and query spacies for the attention probabilities may focus on certain subspaces more than others, potentially leading to more efficient convergence, yet perhaps the Hopfield benefit of exponentially fast convergence may still apply regardless. In contrast, the value weights combine with the multihead attention linear transformation and the fully connected layer, leading to a much more nonconvex problem and training instability.

Dot product approximations for exponential dot product attention have been used, and perhaps those works have some connection to the original Hopfield network Ising model interpretation. Recent work () has further explored the connection between Hopfield networks and Restricted Boltzmann Machines. Perhaps the FC layer or linear attention can be seen as original Hopfield network layers and exponential dot product attention modern Hopfield network layers.

The Hopfield theory seems to attempt continuous integration of states, which requires a quadratic term in the energy to keep the partition function finite. The quadratic term also seems to fulfill another duty of modeling the skip connection. While we perform a discrete normalization, it is reasonable to assume that our set of hidden states are equivalent to hidden states from a continuous distribution. What this distribution is unclear, though it draws parallels to Arora's random walk of language around a context vector, which may be equivalent to McBal's mean-field approximation interpretation of transformers.