A Probabilistic Interpretation of Transformers
International Conference on Machine Learning (ICML 2021)

Alexander Shim [1]

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.

1. Introduction

Transformers have reached state of the art results in language models, significantly outperforming LSTMs. One conceptual explanation for their greater performance is the ability of attention to utilize long range dependencies, whereas Recurrent Neural Networks are limited to encoding past information within a fixed-size hidden state. What this explanation does not explain is how certain architectural choices of transformers, specifically exponential dot product attention, also somewhat ambiguously referred to as softmax attention, outperforms alternatives.

Exponential dot product attention has been popularized in contrastive learning and metric learning and is used in state of the art semi-supervised contrastive learning models. In language models, an exponential dot product probability is used to model conditional probabilities in Word2Vec as well as in some memory networks.

The successes of transformers has been verified empirically, but little work has focused upon a theoretical framework for transformers. We offer a probabilistic explanation, based off of distributions of the exponential family, for attention and contrastive probabilities. Expressing attention as an exponential family allows us to utilize related theory in statistics, machine learning, and statistical mechanics, offering insightful interpretations of the transformer architecture. We provide proofs for attention updates over several continuous distributions as well.

We also explicitly state the limitations of our theory, noting that the modern Hopfield network interpretation of attention shares many of these limitations. For some of these limitations, we speculate connections between other areas of research which may reconcile the theoretical inconsistencies, motivating directions for future research.

2. Background

2.1. Exponential Dot Product Attention

Word2vec used a skip-gram model to predict neighboring words using a conditional distribution defined by a normalized exponential dot product multinoulli function (Mikolov et al., 2013)

Attention was proposed through normalized exponential alignment functions, often referred to as softmax attention in literature, for Neural Machine Translation (Graves, 2013; Bahdanau et al., 2014), and later work parallelizing computation on sequential data introduced normalized exponential dot product similarity (Parikh et al., 2016) (A Decomposable Attention Model for Natural Language Inference). Neural Turing Machines gated memory updates using normalized exponential cosine similarity, in what is referred to as soft attention (Graves et al., 2014).

Other transformer precursors parallelized attention updates over the entire sequence into a layer and switched to convolution-based attention weights (Kaiser & Sutskever, 2016; Kaiser & Bengio, 2016). The transformer architecture incorporated exponential dot product attention scaled by dimensionality (Vaswani et al., 2017).

2.2. Contrastive Learning and Metric learning

Noise-Contrastive Estimation (NCE) creates a mixture distribution between real data and noisy data to convert an unsupervised learning problem into a semi-supervised learning problem, modeling the contrastive probabilities as a

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

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

parametrized logistic distribution (Gutmann & Hyvärinen, 2010).

In metric learning, Multidimensional scaling calculates pairwise distances between projected points (Cox & Cox, 2000), which for Euclidean distances are equivalent to calculating covariance matrix terms using dot products (Bishop, 2007). Neighborhood Components Analysis learns a low dimensional linear embedding matrix and models a probability of a neighbor by comparing the exponential negative square distance $e^{-\|Ax_i - Ax_j\|^2}$ of two input data x_i, x_j to the sum of the exponential negative distances to non-neighbors (Goldberger et al., 2004)

Due to slow convergence of Bernoulli contrastive loss and triplet loss, Sohn proposed an exponential dot product probability over multiple examples (Sohn, 2016), which is mathematically consistent with NCE for multiple distributions. The papers roots in metric learning motivated the dot product form, with a direct influence from Neighborhood Component Analysis.

More recent contrastive learning research adopted a contrastive loss based off of exponential dot product probabilities, including papers that achieve state of the art semi-supervised learning (Wu et al., 2018; Chen et al., 2020).

2.3. Shortcut Connections and Dynamical Systems

Long Short-Term Memory (LSTM) combined a shortcut connection to deal with the vanishing and exploding gradient problem along with gating functions to incorporate and forget information (Hochreiter & Schmidhuber, 1997). Residual connections similarly formulated the hidden layer as an update to an identity mapping, though without a gating mechanism (He et al., 2015). Recurrent Neural Networks have been interpreted as a discrete time approximation to a continuous dynamical system (Jaeger, 2001), where gating acts as a warping of time (Tallec & Ollivier, 2018). Residual connections have been interpreted as a discretized update to a differential equation (Weinan, 2017; Lu et al., 2020).

Interpreting residual networks as discretized differential equations, researchers have posed alternative methods for performing forward updates to converge to equilibrium points and backwards updates to the parameters from the equilibrium points (Chen et al., 2019; Bai et al., 2019). Further work has used monotone operator theory in convex analysis for solving for equilibrium points, interpreting layers as an operator (Winston & Kolter, 2021).

Mathematically similar to our work, transformers have been interpreted as an update of modern hopfield networks and fixed points have been calculated with respect to a fixed set of patterns (Ramsauer et al., 2020). Our work similarly views the attention sublayer as an operator update over a class of discretized probability distribution, though with a changing set of patterns.

2.4. Log Normalizer and Free Energy

Partition functions, or the normalizer function, in statistical physics defines a normalization factor of the Hamiltonian with respect to a parameter defining the temperature. The Boltzmann distribution can be derived through Lagrange multipliers as the distribution which maximizes entropy with a conservation of energy constraint. Jaynes adapted the Boltzmann distributions to maximum entropy distributions with multiple expected statistics constraints by converting the maximum entropy problem into the dual problem of optimizing the log normalizer (Jaynes, 1982), which is known in statistical mechanics as free energy.

Variational methods have been used to approximate log probabilities of observations in machine learning, borrowing from ideas in statistical mechanics. By viewing the joint as an unnormalized probability distribution, the log normalizer is known as the evidence lower-bound, and it has connections to Helmholtz Free Energy (Hinton et al., 1995; Koller & Friedman, 2009).

The sum of exponents loss of AdaBoost (Collins et al.) has been interpreted as the dual form of generalized KL divergence. The log sum of exponents is well known in convex optimization to be the dual form to the maximum entropy objective for a discrete probability distribution (Boyd & Vandenberghe, 2004). Notably, in the modern Hopfield network interpretation of transformers the log sum of exponents is used as part of the energy function and solved through convex optimization techniques (Ramsauer et al., 2020).

In the work most similar to ours, McBal suggested an energy-based model interpretation of transformers, with separate energy updates for each attention sublayer and fully-connected layer, as well as additional ideas from statistical physics (Bal, 2020).

3. Exponential Dot Product Attention

3.1. Exponential Families

The natural parameter form, also known as the canonical form or a linear exponential family, of a distribution of the exponential family can be written as

$$p(x|\eta) = \frac{1}{Z(\eta)} h(x) e^{u(x)^\intercal \eta} \tag{1}$$

where x is the random variable, $u(x)$ the sufficient statistic, η the natural parameter, $h(x)$ the intrinsic measure or carrier measure, and $Z(\eta) := \int h(x) e^{u(x)^\intercal \eta}$ the normalizer or partition function. An exponential family distribution can be generated from an intrinsic measure $h(x)$ by defining the sufficient statistic $u(x) := x$, and defining the normalizer as the Laplace transform applied to the intrinsic measure:

$$Z(\eta) = \mathcal{L} \circ h(x) = \int h(x) e^{x^\intercal \eta} dx \tag{2}$$

When $h(x)$ is chosen to be a uniform discrete measure over a finite set of points $\{x_n\}_{n=1}^N$ in a continuous space, the probability density converts to a probability mass function, and the normalizer is the summation $Z(\eta) = \sum_{n=1}^N e^{u(x_n)^\intercal \eta}$ instead of an integral.

When the query q_i is chosen as the natural parameter and the keys as the finite set of points $\{k_j\}_{j=1}^N$, exponential dot product attention weights are of the form of an exponential family

$$p(k_j|q_i) = \frac{e^{k_j \cdot q_i}}{Z(q_i)} \tag{3}$$

The expected sufficient statistic of an exponential family can be written as a one-to-one function of the natural parameter

$$E_{x \sim P(x|\eta)} \left[u(x) \right] = \nabla_\eta \log Z(\eta) \tag{4}$$

For the exponential family defined by attention, we observe that attention averaging is exactly the gradient of the log normalizer

We further prove results for attention updates of non-discrete $h(x)$. Proofs and further results are given in the appendix.

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 $\nabla_\eta \log \int h(x) e^{x^\intercal \eta} = E_{x \sim p_0(x|\eta_1 + \eta, \eta_2)} \left[x \right]$

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

When the attention sublayer is added to the residual connection, we observe that we are performing a gradient update of the log normalizer with respect to the natural parameters, which are the hidden states.

The log normalizer for this discrete distribution is the log sum of exponents, which is a component of the modern Hopfield energy function, where the attention sublayer also acts as an update for the hidden states

$$\log Z(q_i) = \log \sum_{j=1}^N e^{k_j \cdot q_i} \tag{5}$$

3.2. Log normalizer of exponential families

The set of queries can be seen as IID samples from a distribution of natural parameters [1], meaning our gradient operator is updating a distribution of distributions. In information geometry, natural parameters are naturally mapped to their exponential family distributions, specifically when the Bregman divergence $B_G(\eta_1, \eta_2)$ is defined by the log normalizer $G(\eta) = \log Z(\eta)$, equaling $KL\left(p(x|\eta_1) \| p(x|\eta_2)\right)$.

The Bregman divergence of the convex conjugate of the log normalizer is also equivalent to KL divergence, and the dual variable is exactly the expected sufficient statistic $\bar{u}$, relating back to the activation function (4).

$$\begin{aligned} G^*(\eta^*) &= \sup_\eta \left(\eta^\intercal \eta^* - G(\eta) \right) \\ \eta^* &= \bar{u} = \nabla_\eta \log Z(\eta) \end{aligned} \tag{6}$$

$G^*(\bar{u}) = -H(p(x|\bar{u}) - E_{x \sim p(x|\bar{u})} \left[\log h(x) \right]$ is an offset of the negative entropy function acting on the exponential family distribution with expected sufficient statistic $\bar{u}$, meaning the dual problem of attention is a variant of the maximum entropy problem. More importantly, the duality between natural parameters and expected sufficient statistics can be used to transform a distribution of natural parameters, that is, our query space, into a distribution of expected sufficient statistics, that is, our key space, through our activation function. Unfortunately, the activation function is not always an affine function, but if we attempt an affine approximation, we recover $\nabla^2 \log Z(\eta) = E_{x \sim p(x)} \left[u(x) u(x)^\intercal \right]$, the Fisher information matrix.

When our hidden states is our natural parameter distribution, our affine approximation for the key corresponding to h_j is $\Sigma^{-1} h_j$, exponential dot product attention is proportional to $e^{h_i^T \Sigma^{-1} h_j}$, suggesting our key and query weights $W_q^\intercal W_k \approx \Sigma^{-1}$.

3.3. Expansion and contraction

The attention sublayer outputs a convex combination of the keys, $\sum_{n=1}^N A(k_n, q) k_n$. Without a residual connection, repeated applications of attention would contract the hidden states into the interior, intuitively risking a collapse to a single fixed point.

With a residual connection, assuming our hidden states are recentered around the origin through layer normalization or some other normalization, we could intuitively imagine that roughly radially symmetric hidden states would push each hidden state further away from the origin. Since the log normalizer is the dual form of the maximum entropy distribution, our log normalizer ascent should result in increased

[1] McBal refers to attention as a mean-field approximation

entropy, resulting in an expansion of the hidden states away from their starting points.

We prove that at equilibrium this result holds for uncentered Gaussian distributions. The more formal statement and proof are given in the appendix.

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 $h(x) : \eta \to \Sigma\eta$.*

4. Operator Interpretation

Starting with a hidden state of N tokens, our attention heads are simultaneous gradient updates on each token. We can interpret the tokens as a discretization of a distribution, drawing parallels to the empirical distribution defined by sampling, with the attention operator acting as a pointwise gradient update of the distribution of natural parameters. Two interesting cases of deriving an exponential family from an intrinsic measure are when a conjugate prior $p(\eta|\xi, \nu)$ is defined as the intrinsic measure, and in boosting where the previous model defines an intrinsic measure of unnormalized weights by $w_i = e^{-t_i \eta_n(x_i)}$, where η_n is the previous iteration of the model logits (Collins et al.).

The FC layer can be viewed as a discrete approximation to an operator as well, so the entire attention block can be viewed as a distributional operator. If different attention blocks have tied weights, then it is the same operator applied repeatedly, otherwise the operators for each layer change.

5. Theoretical Limitations

Both this work and the Hopfield interpretation of transformers initially ignore the weight matrices, which are the only parameters that the network learns for the attention sublayer. With transformer models having hundreds of millions to a billion parameters and being difficult to train, this oversight is deeply problematic. The simplest interpretation is for the weight matrices to represent different distortions of space, in which case different attention heads would represent different distortions. Unfortunately, this makes multihead attention incoherent, as we would be combining gradient updates from different spatial transformations. We present a somewhat more consistent lower bound approximation of the log normalizer of a mixture of Gaussians in the appendix.

For this work, we have interpreted a single update on a single hidden state as a gradient update. However, it is less clear why each hidden state has a separate gradient update, and since they depend on the other hidden states, they are not independent updates. The Ising model, which motivated the original Hopfield network, has similar properties. Sample-based approximate inference techniques for Sequential Monte Carlo, particularly ones utilizing bootstrapping, are dependent on the samples and qualitatively similarly use N samples to generate N new samples for the next time step.

The FC layer is inconsistent with both the Hopfield theory of attention and this paper's gradient update of the log normalizer. We would have to view it as a separate operator update composed on top of our attention operator.

6. Future Work

The Hopfield theory of transformers solved for equilibrium behavior, but specifically when the patterns were fixed, whereas for transformers the patterns are the dynamically changing hidden states. A useful experiment would be to sample from an initial distribution and observe how distributions change with each layer, especially with learned weights.

Since the theoretical interpretations ignore the FC layer, it would be useful to know how close the FC layer is to an identity mapping or a purely linear layer, if at all.

Assuming layer normalization should induce an equilibrium distribution of a Gaussian distribution, we could potentially try other contractive mappings to see if this generates substantially different embeddings and layer behavior.

One key difference in theories is that in this work, the residual connection is used as an initial point for a gradient update, whereas the Hopfield theory of attention defines a quadratic term in the Hopfield energy, strongly suggesting that we are working with Gaussian distributions. Since exponential dot product probabilities may come from locally Euclidean assumptions in metric learning, weighted attention may be making inherent assumptions about Gaussian distributions or mixtures of Gaussians over the key space.

Non-Gaussian exponential families will not map the natural parameter to the expected sufficient statistic through an affine transformation. In those cases, we would need to apply a non-linear activation function to transform the natural parameter space into the sufficient statistic space before we take the dot product. An affine approximation of the activation function is the Fisher information matrix, so we could test to see if the query and key weight matrices combine to an approximation of the Fisher information matrix.

References

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate, 2014.

Bai, S., Kolter, J. Z., and Koltun, V. Deep equilibrium models, 2019.

Bal, M. An energy-based perspective on attention mechanisms in transformers, Dec 2020. URL https://mcbal.github.io/post/an-energy-based-perspective-on-attention-mechanisms-in-transformers/.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007. ISBN 0387310738.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 0521833787.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. Neural ordinary differential equations, 2019.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

Collins, M., Schapire, R. E., and Singer, Y. Logistic regression, adaboost and bregman distances. In *Computational Learing Theory*, pp. 158–169.

Cox, T. F. and Cox, M. *Multidimensional Scaling, Second Edition*. Chapman and Hall/CRC, 2 edition, 2000. ISBN 1584880945.

Goldberger, J., Roweis, S. T., Hinton, G. E., and Salakhutdinov, R. Neighbourhood components analysis. In *NIPS*, pp. 513–520, 2004.

Graves, A. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850, 2013.

Graves, A., Wayne, G., and Danihelka, I. Neural turing machines. *CoRR*, abs/1410.5401, 2014.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Teh, Y. and Titterington, M. (eds.), *Proc. Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, volume 9 of *JMLR W&CP*, pp. 297–304, 2010.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. The wake-sleep algorithm for unsupervised neural networks. *Science*, 268:1158–1161, 1995.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Jaeger, H. The echo state approach to analysing and training recurrent neural networks. *GMD-Report 148, German National Research Institute for Computer Science*, 01 2001.

Jaynes, E. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982. doi: 10.1109/PROC.1982.12425.

Kaiser, L. and Bengio, S. Can active memory replace attention? *CoRR*, abs/1610.08613, 2016.

Kaiser, L. and Sutskever, I. Neural gpus learn algorithms, 2016.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press, 2009. ISBN 9780262013192.

Lu, Y., Zhong, A., Li, Q., and Dong, B. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations, 2020.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.

Parikh, A. P., Täckström, O., Das, D., and Uszkoreit, J. A decomposable attention model for natural language inference. *CoRR*, abs/1606.01933, 2016.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Pavlovic, M., Sandve, G. K., Greiff, V., Kreil, D. P., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S. Hopfield networks is all you need. *CoRR*, abs/2008.02217, 2020.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016.

Tallec, C. and Ollivier, Y. Can recurrent neural networks warp time?, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Weinan, E. A proposal on machine learning via dynamical systems. 2017.

Winston, E. and Kolter, J. Z. Monotone operator equilibrium networks, 2021.

Wu, Z., Xiong, Y., Yu, S., and Lin, D. Unsupervised feature learning via non-parametric instance-level discrimination, 2018.