Attention as Implicit Structural Inference

Ryan Singh [1] **Christopher L. Buckley** [1]

Abstract

Attention mechanisms play a crucial role in cognitive systems by allowing them to flexibly allocate cognitive resources. Transformers, in particular, have become a dominant architecture in machine learning, with attention as their central innovation. However, the underlying intuition and formalism of attention in Transformers is based on ideas of keys and queries in database management systems. In this work, we pursue a structural inference perspective, building upon, and bringing together, previous theoretical descriptions of attention such as; Gaussian Mixture Models, alignment mechanisms and Hopfield Networks. Specifically, we demonstrate that attention can be viewed as inference over an implicitly defined set of possible adjacency structures in a graphical model, revealing the generality of such a mechanism. This perspective unifies different attentional architectures in machine learning and suggests potential modifications and generalizations of attention. We hope by providing a new lens on attention architectures our work can guide the development of new and improved attentional mechanisms.

1. Introduction

This depth and breadth of success of the transformer architecture indicates the attention mechanism expresses a useful computational primitive. Recent work has shown interesting theoretical links to kernel methods (Chen et al., 2021; Tsai et al., 2019; Han et al., 2022), Hopfield networks (Ramsauer et al., 2021), and Gaussian mixture models (Li et al., 2019; Movellan & Gabbur, 2020; Gabbur et al., 2021; Shim, 2022; Nguyen et al., 2022), however a formal understanding that captures the generality of this computation remains outstanding. In this paper, we show the attention mechanism can naturally be described as inference on the structure of a graphical model, agreeing with observations that transformers are able to flexibly choose between models based on context (von Oswald et al., 2022; Garg et al.). This Bayesian perspective complements previous theory (Kim et al., 2017; Ramsauer et al., 2021; Shim, 2022), while adding new methods for reasoning about inductive biases and the functional role of attention variables.

This paper proceeds in three parts. First in Sec.3, we show that 'soft' attention mechanisms (e.g. self-attention, cross-attention, graph attention, which we call *transformer attention* hereafter) can be understood as taking an expectation over possible connectivity structures, providing an interesting link between softmax-based attention and marginal likelihood. Second in Sec.4, we extend the inference over connectivity to a Bayesian setting which, in turn, provides a theoretical grounding for iterative attention mechanisms (slot-attention and block-slot attention) (Locatello et al., 2020; Singh et al., 2022), Modern Continuous Hopfield Networks (Ramsauer et al., 2021) and Predictive Coding Networks. Finally in Sec.5, we leverage the generality of this description in order to design new mechanisms with predictable properties.

$$
\begin{aligned}
Attention(Q,K,V) &= \overbrace{softmax(\frac{QW_QW_K^TK^T}{\sqrt{d_k}})}^{p(E\mid Q,K)}W_VV \\
&= \mathbb{E}_{p(E|Q,K)}[V]
\end{aligned}
$$

A key observation is that the attention matrix can be seen as the posterior distribution over edges E, in a graph consisting of query nodes Q and key nodes K. Where the full mechanism computes an expectation of a function defined on the graph $V : \mathcal{G} = (K \cup Q, E) \rightarrow \mathbb{R}^{d\times|\mathcal{G}|}$ with respect to this posterior. This formalism provides an alternate Bayesian theoretical framing within which to understand attention models, shifting the explanation from one centred around retrieval to one that is fundamentally concerned with in-context inference of probabilistic relationships (including retrieval). Within this framework different attention architectures can be described by considering different implicit probabilistic models, by making these explicit we hope to support more effective analysis and the development of new architectures.

[1]School of Engineering and Informatics, University of Sussex, Brighton, United Kingdom. Correspondence to: Ryan Singh <rs773@sussex.ac.uk>.

Accepted to ICML workshop on Structured Probabilistic Inference & Generative Modeling, Honolulu, Hawaii, USA. 2023.

2. Related Work

A key benefit of the perspective outlined here is to tie together different approaches taken in the literature. Specifically, structural variables can be seen as the alignment variables discussed in previous Bayesian descriptions (Kim et al., 2017; Deng et al., 2018; Fan et al., 2020), on the other hand Gaussian Mixture Models (GMMs) (Li et al., 2019; Gabbur et al., 2021; Shim, 2022; Nguyen et al., 2022) can be seen as a specific instance of the framework developed here. This description maintains the explanatory power of GMMs by constraining the alignment variables to be the edges of an implicit graphical model, while offering the increased flexibility of alignment approaches to describe multiple forms of attention (full discussion in Appendix A).

3. Transformer Attention

3.1. Attention as Expectation

We begin by demonstrating transformer attention can be seen as calculating an expectation over graph structures. Specifically, let $x = (x_1, .., x_n)$ be observed input variables, ϕ be some set of discrete latent variables representing edges in a graphical model of x given by $p(x \mid \phi)$, and y a variable we need to predict. Our goal is to find $\mathbb{E}_{y|x}[y]$, however the graph structure ϕ is unobserved so we calculate the marginal likelihood.

$$\mathbb{E}_{y|x}[y] = \sum_{\phi} p(\phi \mid x)\mathbb{E}_{y|x,\phi}[y]$$

Importantly, the softmax function is a natural representation for the posterior $p(\phi \mid x) = softmax(\ln p(x, \phi))$. In order to expose the link to transformer attention, let the model of y given the graph (x, ϕ) be parameterised by a function $\mathbb{E}_{y|x,\phi}[y] = v(x, \phi)$.

$$\mathbb{E}_{y|x}[y] = \sum_{\phi} softmax(\ln p(x, \phi))v(x, \phi) = \mathbb{E}_{\phi|x}[v(x, \phi)] \tag{1}$$

In general, transformer attention can be seen as weighting $v(x, \phi)$ by the posterior distribution $p(\phi \mid x)$ over different graph structures. We show Eq.1 is exactly the equation underlying self and cross-attention by presenting the specific generative models corresponding to them. In this description the latent variables ϕ are identified as edges between observed variables x (keys and queries) in a pairwise Markov Random Field, parameterised by matrices W_K and W_Q, while the function v is parameterised by W_V.

Pairwise Markov Random Fields Given a set of random variables $X = (X_v)_{v \in V}$ with probability distribution $[p]$ and a graph $G = (V, E)$. The variables form a pairwise Markov Random Field (pMRF) (Wainwright & Jordan, 2008) with respect to G if the joint density function $P(X = x) = p(x)$ factorises as follows

$$p(x \mid E) = \frac{1}{Z} \exp\left(\sum_{v \in V} \psi_v + \sum_{e \in E} \psi_e\right)$$

where Z is the partition function $\psi_v(x_v)$ and $\psi_e = \psi_{u,v}(x_u, x_v)$ are known as the node and edge potentials respectively. Beyond the typical set-up, we add a structural prior $p(\phi)$ over a space of possible adjacency structures, $\phi \in \Phi$, of the underlying graph: $p(x, \phi) = p(x \mid \phi)p(\phi)$.

We briefly remark that Eq.1 respects factorisation of $[p]$ in the following sense; if the distribution admits a factorisation (a partition of the space of graphs $\Phi = \prod_i \Phi_i$) with respect to the latent variables $p(x, \phi) = \prod_i f_i(x, \phi_i)$, and the value function distributes over the same partition of edges $v(x, \phi) = \sum_i v_i(x, \phi_i)$ then each of the factors can be marginalised independently:

$$\mathbb{E}_{\phi|x}[v(x, \phi)] = \sum_i \mathbb{E}_{\phi_i|x}[v_i] \tag{2}$$

To recover cross-attention and self-attention we need to specify the structural prior, potential functions and a value function. (In order to ease notation, when Φ_i is a set of edges involving a common node x_i, such that $\phi_i = (x_i, x_j)$ represents a single edge, we use the notation $\phi_i = [j]$, suppressing the shared index.)

3.2. Cross Attention and Self Attention

Key nodes: $K = (x_1, .., x_n)$ and *query nodes*: $Q = (x'_1, ..., x'_m)$. *Structural prior* $p(\phi) = \prod_{i=1}^{m} p(\phi_i)$, where $\Phi_i = \{(x_1, x'_i), .., (x_n, x'_i)\}$ is the set of edges involving x'_i and $\phi_i \sim Uniform(\Phi_i)$ such that each query node is uniformly likely to connect to each key node. *Edge potentials* $\psi(x_j, x'_i) = x_i'^T W_Q^T W_K x_j$, in effect measuring the similarity of x_j and x'_i in a projected space. *Value functions* $v_i(x, \phi_i = [j]) = W_V x_j$, a linear transformation applied to the node at the start of the edge ϕ_i. Taking the posterior expectation in each of the factors defined in Eq.2 gives the standard cross-attention mechanism

$$\mathbb{E}_{p(\phi_i|Q,K)}[v_i] = \sum_j softmax_j(x_i'^T W_Q^T W_K x_j) W_V x_j$$

In contrast self-attention can be derived by considering a similar pMRF with $K = Q$ and directed edges (see Figure 1 and Appendix C.1).

4. Iterative Attention

We continue by extending attention to a latent variable setting, where not all the nodes are observed. In essence applying the attention trick, i.e., a marginalisation of structural variables, to a variational free energy (Evidence Lower Bound).

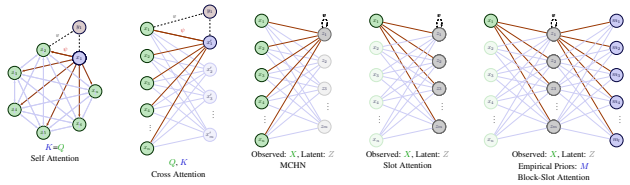

Figure 1. Comparison of models involved in different attention mechanisms. In each case, the highlighted edges indicate Φ_i the support of the uniform prior over ϕ_i.

4.1. Collapsed Inference

We present a version of collapsed variational inference (Teh et al., 2006), where the collapsed variables ϕ are again structural, showing how this results in a Bayesian attention mechanism. In contrast to the previous section, we have a set of (non-structural) latent variables z. The goal is to infer z given the observed variables, x, and a latent variable model $p(x, z, \phi)$. Collapsed inference proceeds by marginalising out the extraneous latent variables ϕ (Teh et al., 2006):

$$p(x, z) = \sum_{\phi} p(x, z, \phi) \tag{3}$$

We define a mean-field recognition density $q(z) \sim \prod_i N(z_i; \mu_i, \Sigma_i)$ and optimise the variational free energy $\min_\lambda \mathbb{E}_q[\ln q_\lambda(z) - \ln p(x, z)]$ with respect to the parameters, $\lambda = (\mu, \Sigma)$, of this distribution. Under a first order laplace approximation, the variational free energy can be expressed as a negative log-likelihood $F \approx -\ln p(x, \mu)$ (Appendix B.1). Substituting in Eq.3 and taking the derivative with respect to the variational parameters:

$$\frac{\partial F}{\partial \mu} = -\frac{1}{\sum_\phi p(x, \mu, \phi)} \sum_{\phi} \frac{\partial}{\partial \mu} p(x, \mu, \phi) \tag{4}$$

In order to make the link to attention, we employ the log-derivative trick, substituting $p_\theta = e^{\ln p_\theta}$ and re-express Eq.4 in two ways:

$$\frac{\partial F}{\partial \mu} = -\sum_{\phi} softmax_\phi(\ln p(x, \mu, \phi)) \frac{\partial}{\partial \mu} \ln p(x, \mu, \phi) \tag{5}$$

$$\frac{\partial F}{\partial \mu} = \mathbb{E}_{\phi|x,\mu}[-\frac{\partial}{\partial \mu} \ln p(x, \mu, \phi)] \tag{6}$$

The first form reveals the softmax which is ubiquitous in all attention models. The second, suggests the variational update should be evaluated as the expectation of the typical variational gradient (the term within the square brackets) with respect to the posterior over the parameters represented by the random variable ϕ. In other words, Bayesian attention is exactly transformer attention applied iteratively where the value function is the variational free energy gradient. We continue by deriving updates for a general pMRF.

Free Energy of a marginalised pMRF Recall the factorised pMRF, $p(x, \phi) = \frac{1}{Z} \prod_i f_i(x, \phi_i)$. Again, independence properties simplify the calculation, the marginalisation can be expressed as a product of local marginals, $\sum_\phi p(x, \phi) = \frac{1}{Z} \prod_i \sum_{\phi_i} f_i(x, \phi_i)$. Returning to the inference setting, the nodes are partitioned into observed nodes, x, and variational parameters μ. Hence the collapsed variational free energy Eq.4, can be expressed as, $F(x, \mu) = -\sum_i \ln \sum_{\phi_i} f_i(x, \mu, \phi_i)$ and it's derivative:

$$\frac{\partial F}{\partial \mu_j} = -\sum_i \sum_{\phi_i} softmax(f_i(x, \mu, \phi_i)) \frac{\partial f_i}{\partial \mu_j}$$

Finally, we follow (Ramsauer et al., 2021) in using the Convex-Concave Procedure (CCCP) to derive a simple fixed point equation which necessarily reduces the free energy (details in Appendix B.1),

$$\mu_j^* = \sum_i \sum_{\phi_i} softmax(g_i(x, \mu, \phi_i)) \frac{\partial g_i}{\partial \mu_j} \tag{7}$$

where $g_i = \sum_{e \in \phi_i} \psi_e$. We follow Sec.3 in specifying specific structural priors and potential functions that recover different iterative attention mechanisms.

4.2. Hopfield-Style Cross Attention

Let the observed, or memory, nodes $x = (x_1, .., x_n)$ and latent nodes $z = (z_1, .., z_m)$ have the following structural prior $p(\phi) = \prod_{i=1}^m p(\phi_i)$, where $\phi_i \sim Uniform\{(x_1, z_i), .., (x_n, z_i)\}$, meaning each latent nodes is uniformly likely to connect to a memory node. Define edge potentials $\psi(x_j, z_i) = z_i W_Q^T W_K x_j$. Application of Eq.7

$$\mu_i^* = \sum_j softmax_j(\mu_i W_Q^T W_K x_j) W_Q^T W_K x_j$$

When μ_i is initialised to some query ξ the system the fixed point update is given by $\mu_i^*(\xi) = \mathbb{E}_{\phi_i|x,\xi}[W_Q^T W_K x_j]$. If the patterns x are well separated, $\mu_i^*(\xi) \approx W_Q^T W_K x_{j'}$, where $W_Q^T W_K x_{j'}$ is the closest vector and hence can be used as an associative memory.

4.3. Slot Attention

Slot attention (Locatello et al., 2020) is an object centric learning module centred around an iterative attention mechanism. Here we show this is a simple adjustment of the prior beliefs on our edge set. With the same set of nodes and potentials, let $p(\phi) = \prod_{j=1}^n p(\phi_j)$, $\phi_j \sim Uniform\{(x_j, z_1), .., (x_j, z_m)\}$. Notice, in comparison to MCHN, the prior over edges is swapped, each observed node is uniformly likely to connect to a latent node, in turn altering the index of the softmax.

$$\mu_i^* = \sum_j softmax_i(\mu_i W_Q^T W_K x_j) W_Q^T W_K x_j$$

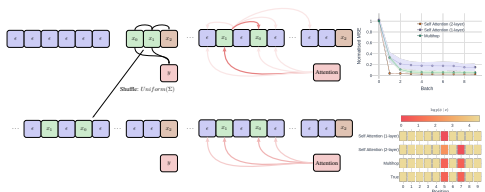

Figure 2. Multihop Attention: (left) Graphical description of the toy problem, x_2 is generated causally from x_1 and x_0, which are used to generate y. (centre) Comparison of the attention employed by Multihop which takes two steps on the attention graph (top) contrasted with Self Attention (bottom). Multihop Attention has the correct bias to learn the task approaching the performance of two-layer Self Attention, while a single layer of Self Attention is unable (top right). Empirically examining the attention weights, Multihop Attention is able to balance attention across two positions, while self-attention favours a single position.

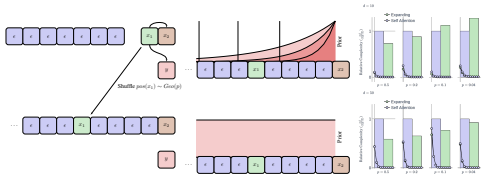

Figure 3. Expanding Attention: (left) Graphical description of the toy problem, x_2 and y are generated from x_1 which is shuffled with a (exponentially decaying) recency bias. (centre) Comparison of the geometric prior, with different shades of red representing the iterative refinements during inference, used by Expanding and uniform prior used by Self Attention. (right) The relative number of operations used by Expanding Attention is beneficial when either the recency bias ($1/p$) or the number of feature dimensions (d) is large, training curves (overlaid) across each of these settings remained roughly equivalent.

while the original slot attention employed an RNN to aid the basic update shown here, the important feature is that the softmax is taken over the 'slots'. This forces competition between slots to account for the observed variables, creating object centric representations.

5. New Designs

By identifying the attention mechanism in terms of an implicit probabilistic model, we can review and modify the underlying modelling assumptions in a principled manner to design new attention mechanisms. Recall transformer attention can be written as the marginal probability $p(y \mid x) = \sum_\phi p(\phi \mid x)\mathbb{E}_{y|x,\phi}[y]$, the specific mechanism is therefore informed by three pieces of data: (a) the value function $p(y \mid x, \phi)$, (b) the likelihood $p(x \mid \phi)$ and (c) the prior $p(\phi)$. Here, we explore modifying (a) and (c) and show they can exhibit favourable biases on toy problems.

Multi-hop Attention Our description makes it clear that the value function employed by transformer attention can be extended to any function over the graph. For example, consider the calculation of $\mathbb{E}_{y|x,\phi}[y_i]$ in transformer attention, a linear transformation is applied to the most likely neighbour, x_j, of x_i. A natural extension is to include a two-hop neighbourhood, additionally using the most likely neighbour x_k of x_j. The attention mechanism then takes a different form $\mathbb{E}_{p(\phi_j|\phi_i)p(\phi_i|x)}[V(x_{\phi_i} + x_{\phi_j})] = (P_\phi + P_\phi^2)VX$, where P_ϕ is the typical attention matrix (Appendix B.4). While containing the same number of parameters as a single-layer of transformer attention, for some datasets two-hop attention should be able to approximate the behaviour of two-layers of transformer attention (Figure 2).

Expanding Attention One major limitation of transformer attention is the reliance on a fixed context window. From one direction, a small context window does not represent long range relationships, on the other hand a large window does an unnecessary amount of computation when modelling a short range relationship. By replacing the uniform prior with a geometric distribution $p(\phi \mid q) \sim Geo(q)$, and a conjugate hyper-prior $p(q) \sim Beta(\alpha, \beta)$, and using a (truncated) mean-field variational inference procedure (Zobay, 2009), we derive a mechanism (Appendix B.4) that dynamically scales depending on input (Figure 3).

6. Discussion

In this paper, we presented a probabilistic description of the attention mechanism, formulating attention as structural inference within a probabilistic model. This approach builds upon previous research that connects cross attention to inference in a Gaussian Mixture Model. By considering the discrete inference step in a Gaussian Mixture Model as inference on marginalised structural variables, we bridge the gap with alignment-focused descriptions. This framework naturally extends to self-attention, graph attention, and iterative mechanisms, such as Hopfield Networks additionally allowing us to discuss the relationship to cognitive theories of attention (Appendix B.2).

We hope this work will contribute to a more unified understanding of the functional advantages and disadvantages brought by Transformers. Furthermore, we argue that viewing Transformers from a structural inference perspective provides different insights into their central mechanism. Typically, optimising structure is considered a learning problem, changing on a relatively slow timescale compared to inference. However, understanding Transformers as fast structural inference suggests that their remarkable success stems from their ability to change effective connectivity on the same timescale as inference.

This general idea can potentially be applied to various architectures and systems. For instance, Transformers employ relatively simple switches in connectivity compared to the

complex dynamics observed in the brain (Tognoli & Kelso, 2014). Exploring inference over more intricate structural distributions, such as connectivity motifs or modules in network architecture, could offer artificial systems even more flexible control of resources.

6.1. Limitations and Future Directions

The connection to structural inference presented here is limited to the attention computation of a single transformer head. While positional encodings are naturally incorporated (see appendix for details), an interesting future direction would be to investigate whether multiple layers and multiple heads typically used in a transformer block can also be interpreted within this framework. Additionally, the extension to iterative inference employed a crude approximation to the variational free energy, arguably destroying the favourable properties of Bayesian methods. Suggesting the possibility of creating iterative attention mechanisms with second order approximation terms, possibly producing more robust mechanisms.

Acknowledgements

This work was supported by The Leverhulme Trust through the be.AI Doctoral Scholarship Programme in biomimetic embodied AI. Additional thanks to Alec Tschantz, Tomasso Salvatori, Miguel Aguilera and Tomasz Korbak for their invaluable feedback and discussions.

References

Annabi, L., Pitti, A., and Quoy, M. On the Relationship Between Variational Inference and Auto-Associative Memory, October 2022.

Buckley, C. L., Kim, C. S., McGregor, S., and Seth, A. K. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79, December 2017. ISSN 0022-2496. doi: 10.1016/j.jmp.2017.09.004.

Carandini, M. and Heeger, D. J. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, January 2012. ISSN 1471-0048. doi: 10.1038/nrn3136.

Chen, Y., Zeng, Q., Ji, H., and Yang, Y. Skyformer: Remodel Self-Attention with Gaussian Kernel and Nystr\"om Method, October 2021.

Clark, A. The many faces of precision (Replies to commentaries on "Whatever next? Neural prediction, situated agents, and the future of cognitive science"). *Frontiers in Psychology*, 4, 2013. ISSN 1664-1078.

Deng, Y., Kim, Y., Chiu, J., Guo, D., and Rush, A. Latent Alignment and Variational Attention. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Fan, X., Zhang, S., Chen, B., and Zhou, M. Bayesian Attention Modules. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16362–16376. Curran Associates, Inc., 2020.

Feldman, H. and Friston, K. Attention, Uncertainty, and Free-Energy. *Frontiers in Human Neuroscience*, 4, 2010. ISSN 1662-5161.

Frecon, J., Gasso, G., Pontil, M., and Salzo, S. Bregman Neural Networks. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 6779–6792. PMLR, June 2022.

Friston, K. and Kiebel, S. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1211–1221, May 2009. ISSN 0962-8436. doi: 10.1098/rstb.2008.0300.

Gabbur, P., Bilkhu, M., and Movellan, J. Probabilistic Attention for Interactive Segmentation, July 2021.

Garg, S., Tsipras, D., Liang, P., and Valiant, G. What Can Transformers Learn In-Context? A Case Study of Simple Function Classes.

Han, X., Ren, T., Nguyen, T. M., Nguyen, K., Ghosh, J., and Ho, N. Robustify Transformers with Robust Kernel Density Estimation, October 2022.

Kim, Y., Denton, C., Hoang, L., and Rush, A. M. Structured Attention Networks, February 2017.

Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., and Liu, H. Expectation-Maximization Attention Networks for Semantic Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9166–9175, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-72814-803-8. doi: 10.1109/ICCV.2019.00926.

Lindsay, G. W. Attention in Psychology, Neuroscience, and Machine Learning. *Frontiers in Computational Neuroscience*, 14, 2020. ISSN 1662-5188.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., and Kipf, T. Object-Centric Learning with Slot Attention, October 2020.

Millidge, B., Song, Y., Salvatori, T., Lukasiewicz, T., and Bogacz, R. A Theoretical Framework for Inference and Learning in Predictive Coding Networks, August 2022.

Mirza, M. B., Adams, R. A., Friston, K., and Parr, T. Introducing a Bayesian model of selective attention based on active inference. *Scientific Reports*, 9(1):13915, September 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-50138-8.

Movellan, J. R. and Gabbur, P. Probabilistic Transformers, November 2020.

Nguyen, T. M., Nguyen, T. M., Le, D. D. D., Nguyen, D. K., Tran, V.-A., Baraniuk, R., Ho, N., and Osher, S. Improving Transformers with Probabilistic Attention Keys. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16595–16621. PMLR, June 2022.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S. Hopfield Networks is All You Need, April 2021.

Rao, R. P. N. and Ballard, D. H. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2 (1):79–87, January 1999. ISSN 1546-1726. doi: 10.1038/4580.

Reynolds, J. H. and Heeger, D. J. The Normalization Model of Attention. *Neuron*, 61(2):168–185, January 2009. ISSN 08966273. doi: 10.1016/j.neuron.2009.01.002.

Shankar, S., Garg, S., and Sarawagi, S. Surprisingly Easy Hard-Attention for Sequence to Sequence Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 640–645, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1065.

Shim, A. A Probabilistic Interpretation of Transformers, April 2022.

Singh, G., Kim, Y., and Ahn, S. Neural Block-Slot Representations, November 2022.

Teh, Y., Newman, D., and Welling, M. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

Tognoli, E. and Kelso, J. A. S. The Metastable Brain. *Neuron*, 81(1):35–48, January 2014. ISSN 0896-6273. doi: 10.1016/j.neuron.2013.12.022.

Tsai, Y.-H. H., Bai, S., Yamada, M., Morency, L.-P., and Salakhutdinov, R. Transformer Dissection: A Unified Understanding of Transformer's Attention via the Lens of Kernel, November 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention Is All You Need, December 2017.

von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent, December 2022.

Wainwright, M. J. and Jordan, M. I. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, November 2008. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000001.

Yang, Y., Huang, Z., and Wipf, D. Transformers from an Optimization Perspective, May 2022.

Yuille, A. L. and Rangarajan, A. The Concave-Convex Procedure (CCCP). In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

Zobay, O. Mean field inference for the Dirichlet process mixture model. *Electronic Journal of Statistics*, 3(none): 507–545, January 2009. ISSN 1935-7524, 1935-7524. doi: 10.1214/08-EJS339.

Table 1. Different attention modules

Name	Graph (G)	Prior ($p(\phi)$)	Potentials (ψ)	Value $v(x, \phi)$
Cross Attention	Key nodes K, query nodes Q	Uniform	$x_i'^T W_Q^T W_K x_j$	Vx_j
Self Attention	K = Q, directed edges	Uniform	$x_i^T W_Q^T W_K x_j$	Vx_j
Graph Attention, Sparse Attention	K = Q, directed edges	Uniform (restricted)	$x_i^T W_Q^T W_K x_j$	Vx_j
Relative Positional Encodings	K = Q, directed edges	Categorical	$x_i^T W_Q^T W_K x_j$	Vx_j
Absolute Positional Encodings	K = Q	Uniform	$\tilde{x_i}^T W_Q^T W_K \tilde{x_j}$ $\tilde{x_i} = x_i + e_i$	Vx_j
Classification Layer	NN output $f_\theta(X)$, classes y	Uniform	$f_\theta(X)_i^T y_j$	y_j
MCHN	Observed nodes X, latent nodes Z	Uniform (observed)	$z_i^T W_Q^T W_K x_j$	$\frac{\partial F}{\partial z}$
Slot Attention	Observed nodes X latent nodes Z	Uniform (latent)	$z_i^T W_Q^T W_K x_j$	$\frac{\partial F}{\partial z}$
Block-Slot Attention	Observed nodes X, latent nodes Z, memory nodes M	Uniform (latent)	$z_i^T W_Q^T W_K x_j$, $m_k^T W_Q^T W_K z_i$	$\frac{\partial F}{\partial z}$
PCN	Observed nodes X, multiple layers of latent nodes $\{Z^{(l)}\}_{l \leq L}$	Uniform (latent)	$z_i^T W_Q^T W_K x_j$	$\frac{\partial F}{\partial z}$
Multihop Attention	K = Q, directed edges	Uniform	$x_i^T W_Q^T W_K x_j$	$Vx_j + Vx_k$
Expanding Attention	K = Q, directed edges	Geometric x Beta	$x_i^T W_Q^T W_K x_j$	Vx_j

Here we include some more detailed derivations of claims made in the paper, and list the hyperparameters used for the experiments.

A. Related Work

Latent alignment and Bayesian Attention Several attempts have been made to combine the benefits of soft (differentiability) and stochastic attention, often viewing attention as a probabilistic alignment problem. Most approaches proceed by sampling, e.g., using the REINFORCE estimator (Deng et al., 2018) or a $topK$ approximation (Shankar et al., 2018). Two notable exceptions are (Kim et al., 2017) which embeds an inference algorithm within the forward pass of a neural network, and (Fan et al., 2020) which employs the re-parameterisation trick for the alignment variables. In this work, rather than treating attention weights as an independent learning problem, we aim to provide the implicit model that would give rise to the attention weights.

Relationship to Gaussian mixture model Previous works that have taken a probabilistic perspective on the attention mechanism note the connection to inference in a gaussian mixture model (Li et al., 2019; Gabbur et al., 2021; Shim, 2022; Nguyen et al., 2022). Indeed (Annabi et al., 2022) directly show the connection between the Hopfield energy and the variational free energy of a Gaussian mixture model. Although Gaussian mixture models, a special case of the framework we present here, are enough to explain cross attention they do not capture slot or self-attention, obscuring the generality underlying attention mechanisms. In contrast, the description presented here extends to structural inductive biases beyond what can be expressed in a Gaussian mixture model.

Attention as bi-level optimisation Mapping feed-forward architecture to a minimisation step on a related energy function has been called unfolded optimisation (Frecon et al., 2022). Taking this perspective can lead to insights about the inductive biases involved for each architecture. It has been shown that the cross-attention mechanism can be viewed as an optimisation step on the energy function of a form of Hopfield Network (Ramsauer et al., 2021), providing a link between attention and associative memory. while (Yang et al., 2022) extend this view to account for self-attention. Our framework distinguishes Hopfield attention, which does not allow an arbritary value matrix, from transformer attention. while there remains a strong theoretical connection, it places the Hopfield Energy as an instance of variational free energy, aligning more closely with iterative attention mechanisms such as slot-attention.

B. Mathematical Details

B.1. Iterative Attention

In this section we provide a more detailed treatment of the Laplace approximation, and provide proper justification for invoking the CCCP. For both, the following lemma is useful:

Lemma B.1. *The function* $\ln p(x) = \ln \sum_\phi p(x,\phi) = \ln \sum_\phi \exp E_\phi(x)$ *has derivatives (i)* $\frac{\partial}{\partial x} \ln p(x) = \mathbb{E}_{\phi|x}[\frac{\partial}{\partial x} E_\phi]$ *and (ii)* $\frac{\partial^2}{\partial x^2} \ln p(x) = Var_{\phi|x}[\frac{\partial}{\partial x} E_\phi] + \mathbb{E}_{\phi|x}[\frac{\partial^2}{\partial x^2} E_\phi]$

Proof. Let $E = (E_\phi)$ the vector of possible energies, and $p = (p_\phi) = (p(\phi \mid x))_\phi$ the vector of conditional probabilities. Consider $\ln p(\phi \mid x)$ written in canonical form,

$$\ln p(\phi \mid x) = \langle E_\phi(x), \mathbb{1}_\phi \rangle - A[E_\phi(x)] + h(\phi)$$

Where $A[E(x)] = \ln Z(E)$ is the cumulant generating function. By well known properties of the cumulant: $\frac{\partial A}{\partial E_i} = p(\phi = i \mid x) = p_i$. Hence by the chain rule for partial derivatives, $\frac{\partial A}{\partial x} = \sum_\phi p(\phi \mid x) \frac{\partial}{\partial x} E_\phi$, which is (i).

To find the second derivative we apply again the chain-rule $\frac{d}{dt} f(g(t)) = f''(g(t))g'(t)^2 + f'(g(t))g''(t)$. Again by properties of the cumulant $\frac{\partial^2 A}{\partial E_i \partial E_j} = Cov(\mathbb{1}_i, \mathbb{1}_j) = [diag(p) - p^T p]_{i,j} = \mathbb{V}_{i,j}$. Hence the second derivative is

$$\frac{\partial^2 A}{\partial^2 x} = \frac{\partial E}{\partial x}^T \mathbb{V} \frac{\partial E}{\partial x} + \mathbb{E}[\frac{\partial^2 E_\phi}{\partial x^2}] \tag{8}$$

□

Second order Laplace Approximation With these derivatives in hand we can calculate the second order laplace approximation of the free energy $\mathcal{F} = \mathbb{E}_q[\ln q_\lambda(z) - \ln p(x,z)]$.

$$\begin{aligned}
\mathcal{F} &\approx \mathbb{E}_q[\ln p(\mu, x) + \frac{\partial}{\partial z} \ln p(\mu, x)^T (z-\mu) + (z-\mu)^T \frac{\partial^2}{\partial z^2} \ln p(\mu, x)(z-\mu)] + H[q] \\
&\approx \ln p(\mu, x) + tr(\Sigma_q^{-1} Var_{\phi|\mu,x}[\frac{\partial}{\partial z} E_\phi]) + tr(\Sigma_q^{-1} \mathbb{E}_{\phi|\mu,x}[\frac{\partial^2}{\partial z^2} E_\phi]) + \frac{1}{2} log \mid \Sigma_q \mid + C
\end{aligned}$$

Optimising this second order approximation w.r.t (μ, Σ) could lead to more robust iterative attention mechanisms. However, the second order terms have both μ and Σ dependence making this approximation difficult to use in practice. As alluded to in the paper, iterative attention mechanisms can also be viewed as an alternating maximisation procedure, which breaks this dependence:

As Alternating Minimisation Collapsed Inference can also be seen as co-ordinate wise variational inference (Teh et al., 2006). Consider the family of distributions $Q = \{q(z;\lambda)q(\phi \mid z)\}$, where $q(z;\lambda)$ is parameterised, however $q(\phi)$ is unconstrained.

$$\begin{aligned}
\mathcal{F} &= \min_{q \in Q} \mathbb{E}_q[\ln q(z,\phi) - \ln p(x,z,\phi)] \\
&= \min_{q \in Q} \mathbb{E}_{q(z)}[\mathbb{E}_{q(\phi)}[\ln q(\phi) - \ln p(x, \phi \mid z)] + \ln q(z) - \ln p(z)]
\end{aligned}$$

The inner expectation is maximised for $q(\phi) = p(\phi \mid x, z)$ and the inner expectation evaluates to $-\ln p(x \mid z)$ which recovers the marginalised objective

$$\min_{q \in Q} \mathbb{E}_{q(z)}[q(z) - \ln \sum_{\phi} p(x, z, \phi)]$$

This motivates an alternate derivation of iterative attention as structural inference which is less reliant on the Laplace approximation; Consider optimising over the variational family $Q = \{q(z; \lambda)q(\phi)\}$ coordinate wise:

$$\begin{aligned} \ln q_{t+1}(\phi) &= \mathbb{E}_{q_t(z;\lambda_t)}[\ln p(\phi \mid x, z)] + C \\ \lambda_{t+1} &= arg\min_{\lambda} \mathbb{E}_{q_t(\phi)}[\mathbb{E}_{q(z;\lambda)}[\ln q(z) - \ln p(x, z \mid \phi)] \\ &= arg\min_{\lambda} \mathbb{E}_{q_t(\phi)}[\mathcal{F}_{\phi}] \end{aligned}$$

In the case of quadratic potentials, $q_{t+1}(\phi) = p(\phi \mid x, \lambda_t)$, hence the combined update step can be written

$$arg\min_{\lambda} \mathbb{E}_{p(\phi|x,\lambda_t)}[\mathcal{F}_{\phi}(\lambda)]$$

Each step necessarily reduces the free energy of the mean-field approximation, so this process converges. This derivation is independent of which approximation or estimation is used to minimise the typical variational free energy.

Quadratic Potentials and the Convex Concave Procedure Assuming the node potentials are quadratic $\psi(x_i) = -\frac{1}{2}x_i^2$ and the edge potentials have the form $\psi(x_i, x_j) = x_i W x_j$, and define $g_i = \sum_{e \in \phi_i} \psi_e$. Consider the following fixed point equation,

$$\mu_j^* = \sum_i \sum_{\phi_i} softmax(g_i(x, \mu, \phi_i)) \frac{\partial g_i}{\partial \mu_j} \tag{9}$$

since node potentials are convex and edge potentials are concave, we can invoke the CCCP (Yuille & Rangarajan, 2001), hence this fixed point equation has the property $F(x, \mu_j^*) \leq F(x, \mu_j)$ with equality if and only if μ_j^* is a stationary point of F.

Convexity details for the CCCP Given a pairwise pMRF with quadratic potentials $\psi(x_i) = -\frac{1}{2}x_i^2$ and the edge potentials have the form $\psi(x_i, x_j) = x_i W x_j$ and W p.s.d., s.t. $\ln p(x, \phi) = -\frac{1}{2}\sum_{v \in \mathcal{G}} x_v^2 + \ln \sum_{\phi} \exp g_{\phi}(x)$, where $g_{\phi}(x) = \sum_{e \in \phi} \psi_e$. We need the following lemma to apply the CCCP:
Lemma B.2. $\ln \sum_{\phi} \exp g_{\phi}(x)$ is convex in x.

Proof. We reapply Lemma.B.1, with $E_{\phi} = g_{\phi}(x)$, hence $\frac{\partial^2}{\partial x^2} \ln \sum_{\phi} \exp g_{\phi}(x) = Var_{\phi|x}[\frac{\partial}{\partial x} g_{\phi}] + \mathbb{E}_{\phi|x}[\frac{\partial^2}{\partial x^2} g_{\phi}]$. The first matrix is a variance, so p.s.d. The second term $\mathbb{E}_{\phi|x}[\sum_{e \in \phi} \frac{\partial^2}{\partial x^2} \psi_e]$ is a convex sum of p.s.d matrices. Hence both terms are p.s.d, implying $\ln \sum_{\phi} \exp g_{\phi}(x)$ is indeed convex. □

B.2. Predictive Coding Networks

Predictive Coding Networks (PCN) have emerged as an influential theory in Computational Neuroscience (Rao & Ballard, 1999; Friston & Kiebel, 2009; Buckley et al., 2017). Building on theories of perception as inference and the Bayesian brain, PCNs perform approximate Bayesian inference by minimising a variational free energy of a graphical model, where incoming sensory data are used as observations. Typical implementations use a hierarchical model with Gaussian conditionals, resulting in a local prediction error minimising scheme. The minimisation happens on two distinct time-scales, which can be seen as E-step and M-steps on the variational free energy: a (fast) inference phase encoded by neural activity corresponding to perception and a (slow) learning phase associated with synaptic plasticity. Gradient descent on the free energy gives the inference dynamics for a particular neuron μ_i, (Millidge et al., 2022)

$$\frac{\partial \mathcal{F}}{\partial \mu_i} = -\sum_{\phi^-} k_{\phi} \epsilon_{\phi} + \sum_{\phi^+} k_{\phi} \epsilon_{\phi} w_{\phi}$$

Where ϵ are prediction errors, w represent synaptic strength, k are node specific precisions representing uncertainty in the generative model and ϕ^-, ϕ^+ represent pre-synaptic and post-synaptic terminals resectively. Applying a uniform prior over

the incoming synapses results in a slightly modified dynamics,

$$\frac{\partial \mathcal{F}}{\partial \mu_i} = -\sum_{\phi^-} softmax(-{\epsilon_\phi}^2)k_\phi \epsilon_\phi + \sum_{\phi^+} softmax(-{\epsilon_\phi}^2)k_\phi \epsilon_\phi w_\phi$$

where the softmax function induces a normalisation across prediction errors received by a neuron. This dovetails with theories of attention as normalisation in Psychology and Neuroscience (Reynolds & Heeger, 2009; Carandini & Heeger, 2012; Lindsay, 2020). In contrast previous predictive coding based theories of attention have focused on the precision terms, k, due to their ability to up and down regulate the impact of prediction errors (Clark, 2013; Feldman & Friston, 2010; Mirza et al., 2019). Here we see the softmax terms play a functionally equivalent role to precision variables, inheriting their ability to account for bottom-up and top-down attention, while exhibiting the fast winner-takes-all dynamics that are associated with cognitive attention.

B.3. PCN Detailed Derivation

Here we go through the derivations for the equations presented in section **??**. PCNs typically assume a hierarchical model with gaussian residuals:

$$\begin{aligned} z_0 &\sim N(\hat{\mu_0}, \Sigma_0) \\ z_{i+1} \mid z_i &\sim N(f_i(z_i; \theta_i), \Sigma_i) \\ y \mid z_N &\sim N(f_N(z_N; \theta_N), \Sigma_N) \end{aligned}$$

Under these conditions, a delta approximation of the variational free energy is given by:

$$\begin{aligned} \mathcal{F}[p, q] &= \mathbb{E}_{q(z;\mu)}[-\ln p(y, z)] + H[q] \\ \mathcal{F}(\mu, \theta) &\approx \sum_{l=0}^{N} \Sigma_l^{-1} \epsilon_l^2 \end{aligned}$$

Where $\epsilon_l = (\mu_{l+1} - f_l(\mu_l; \theta_l))^2$. The inference phase involves adjusting the parameters, μ in the direction of the gradient of $\mathcal{F}$, which for a given layer is:

$$\frac{\partial \mathcal{F}}{\partial \mu_l} = \Sigma_{l-1}^{-1} \epsilon_{l-1} - \Sigma_l^{-1} \epsilon_l f'(\mu_l) \tag{10}$$

Here, for ease of comparison, we consider the case where the link functions are linear, $f_i(\cdot) = W_i(\cdot)$ and further the precision matrices are diagonal $\Sigma_i^{-1} = diag(k_i)$. Under these conditions we can write the derivative component-wise as sums of errors over incoming and outgoing edges :

$$(\frac{\partial \mathcal{F}}{\partial \mu_l})_i = -\sum_{\phi^-} k_\phi \epsilon_\phi + \sum_{\phi^+} k_\phi \epsilon_\phi w_\phi$$

Where ϕ^-, ϕ^+ represent the set of incoming and outgoing edges respectively, and we redefine $\epsilon_\phi = (\mu_i - \mu_j w_{ij})$ for an edge $\phi = (z_i, z_j)$ and $k_\phi = K(z_j)$ the precision associated with the node at the terminus of ϕ.

Now if we instead assume a uniform prior over incoming edges, or concretely;

$$\begin{aligned} z_0 &\sim N(\hat{\mu_0}, \Sigma_0) \\ \phi_l^i &\sim Uniform(\{(z_{l+1}^i, z_l^0), (z_{l+1}^i, z_l^1), ...\} \\ z_{l+1}^i \mid z_l, \phi_l^i &\sim N(w_l^{ij} z_l^{\phi_l^i}, 1/k_l^i) \\ y \mid z_N &\sim N(f_N(z_N; \theta_N), \Sigma_N) \end{aligned}$$

The system becomes a pMRF with edge potentials given by the prediction errors, recall applying Eq.4:

$$\frac{\partial F}{\partial \mu_j} = -\sum_i \sum_{\phi_i} softmax(f_i(x, \mu, \phi_i)) \frac{\partial f_i}{\partial \mu_j}$$

Here for a node in a given layer, it participates in one Φ_{l-1}^{j} and all the Φ_{l+1}^{k} from the layer above, where every $f_i(x, \mu, \phi_i)$ here is a squared prediction error corresponding to the given edge $e_l^{ij} = k_l^{ij}(z_l^i - w_l^{ij} z_{l-1}^j)^2$, hence:

$$\begin{aligned}\frac{\partial F}{\partial \mu_j} =& -\sum_{i \in \Phi_{l-1}^j} softmax_i(-(\epsilon_{l-1}^{ij})^2)\epsilon_{l-1}^{ij} k_j \\ &+ \sum_{k \in [l]} \sum_{i' \in \Phi_l^k} softmax_{i'}(-(\epsilon_l^{i'k})^2)\epsilon_l^{i'k} w_l^{i'k} \mathbb{1}(i' = j) \\ \frac{\partial F}{\partial \mu_j} =& -\sum_{i \in \Phi_{l-1}^j} softmax_i(-(\epsilon_{l-1}^{ij})^2)\epsilon_{l-1}^{ij} \\ &+ \sum_{k \in [l]} softmax_{i'}(-(\epsilon_l^{i'k})^2)\epsilon_l^{i'k} w_l^{i'k}\end{aligned}$$

Here incoming signals (nodes i) compete through the softmax, whilst the outgoing signal competes with other outgoing signals from nodes (nodes i') in the same layer for representation in the next layer (nodes k), see block-slot attention diagram for intuition. By abuse of notation (reindexing edges as ϕ)

$$\frac{\partial \mathcal{F}}{\partial \mu_i} = -\sum_{\phi^-} softmax(-{\epsilon_\phi}^2)k_\phi \epsilon_\phi + \sum_{\phi^+} softmax(-{\epsilon_\phi}^2)k_\phi \epsilon_\phi w_\phi$$

While we derived these equations for individual units to draw an easy comparison to standard Predictive Coding, we note it is likely more useful to consider blocks of units competing with each other for representation, similar to multidimensional token representations in typical attention mechanisms. We also briefly note here, the Hammersley–Clifford theorem indicates a deeper duality between attention as mediated by precision matrices and as structural inference.

B.4. New Designs

Multihop Derivation $\mathbb{E}_{y|x,\phi}[y_i]$ in transformer attention, a linear transformation is applied to the most likely neighbour, x_j, of x_i. A natural extension is to include a two-hop neighbourhood, additionally using the most likely neighbour x_k of x_j. Formally, the value function v no longer neatly distributes over the partition Φ_i, however the attention mechanism then takes a different form: $\mathbb{E}_{p(\phi_j|\phi_i)p(\phi_i|x)}[V(x_{\phi_i} + x_{\phi_j})] = (P_\phi + P_\phi^2)VX$. Where we use $\phi_{j(i)} = \phi_j$ to denote the edge set of the node at the end of ϕ_i. To see this note:

$$\begin{aligned}\mathbb{E}_{p(\phi|x)}[V(x_{\phi_i} + x_{\phi_j})] &= \sum_\phi \prod_k p(\phi_k \mid x) V(x_{\phi_i} + x_{\phi_j}) \\ &= \sum_\phi \prod_k p(\phi_k \mid x) V(x_{\phi_i} + x_{\phi_j}) \\ &= \sum_\phi \prod_k p(\phi_k \mid x) V x_{\phi_i} + \sum_\phi \prod_k p(\phi_k \mid x) V x_{\phi_j} \\ &\text{by independence properties} \\ &= \sum_{\phi_i} p(\phi_i \mid x) V x_{\phi_i} + \sum_{\phi_i, \phi_j} p(\phi_i \mid x) p(\phi_j \mid x) V x_{\phi_j}\end{aligned}$$

Denoting the typical attention matrix, P, where $p_{ij} = p(\phi_i = [j] \mid x)$

$$\begin{aligned}&= \sum_k \sum_j p_{jk} p_{ij} V x_k + \sum_j p_{ij} V x_j \\ &= (P_\phi + P_\phi^2) V X\end{aligned}$$

Expanding We iteratively approximate $p(\phi \mid x)$ using the updates: 1. $q_t \leftarrow \frac{\beta_t}{\alpha_t + \beta_t}$, 2. $p_t = p(\phi \mid x, q_t)$, 3. $\alpha_{t+1} \leftarrow \alpha_t + 1$, $\beta_{t+1} \leftarrow \beta_t + \sum_{<H(q_t)} i(p_t)_i$. Where α and β are hyperparameters determining the strength of the prior and H is the

truncation horizon. Since attention dot products can be cached and reused for each calculation of step 2. the iterative procedure is computationally cheap.

The attention mechanism has asymptotic time complexity $O(n^2 d)$ where n is the size of the size of the context window and d is dimension over which the inner product is computed. In comparison, expanding attention $O(n(md + k))$ where m is the size of the window at convergence, and k is the number of steps to converge. If, as is typical, d is large such that $d >> k$ the time complexity of expanding attention should be favourable.

Expanding Derivation As in the main text, let $p(\phi \mid q) \sim Geo(q)$ and $p(q) \sim Beta(\alpha, \beta)$, such that we have the full model $p(x, \phi, q; \alpha, \beta) = p(x \mid \phi)p(\phi \mid q)p(q; \alpha, \beta)$. In order to find $p(\phi \mid x)$ we employ a truncated Mean Field Variational Bayes (Zobay, 2009), assuming a factorisation $p_t(\phi, q) = p_t(\phi)p_t(q)$, and using the updates:

$$\ln p_{t+1}(\phi) = \mathbb{E}_{p_t(q)}[\ln p(x \mid \phi) + \ln p(\phi \mid q)] + C_1$$
$$\ln p_{t+1}(q) = \mathbb{E}_{p_t(\phi)}[\ln p(\phi \mid q) + \ln p(q; \alpha, \beta)] + C_2$$

By conjugacy the second equation simplifies to a simple update of the beta distribution

$$\begin{aligned} \implies p_{t+1}(q) &= Beta(\alpha_{t+1}, \beta_{t+1}) \\ \alpha_{t+1} &= \alpha_t + 1 \\ \beta_{t+1} &= \beta_t + \mathbb{E}_{p_t(\phi)}[\phi] \end{aligned}$$

While the second update can be seen as calculating the posterior given $q_t = \mathbb{E}_{p_t(q)}[q]$,

$$\begin{aligned} \ln p_{t+1}(\phi) &= \ln p(x \mid \phi) + \mathbb{E}_{p_t(q)}[\ln p(\phi \mid q)] + C_2 \\ &= \ln p(x \mid \phi) + \phi \mathbb{E}_{p_t(q)}[\ln q] + C_2 \\ &= \ln p(\phi \mid x, q_t) \end{aligned}$$

Finally, we use a truncation to approximate the infinite sum $\mathbb{E}_{p_t(\phi)}[\phi] = \sum_k p_t(\phi = k)k \approx \sum_{<H} p_t(\phi = k)k$. Where we set the horizon according to the current distribution of q. For example in our experiments we chose $H(q_t) = \ln 0.05 / \ln(1 - q_t)$ the truncation that would capture 95% of the probability mass of the prior.

C. Attention Variants

Here we briefly discuss some variants of attention that there wasn't space for in the paper.

C.1. Self Attention

Details of the MRF for self-attention:

- Nodes $K = Q = (x_1, .., x_n)$
- Structural prior, over a fully connected, directed graph $p(\overrightarrow{\phi}) = \prod_{i=1}^{n} p(\overrightarrow{\phi}_i)$, where $\overrightarrow{\Phi}_i = \{(x_1, x_i), .., (x_n, x_i)\}$ is the set of edges involving x_i and $\overrightarrow{\phi}_i \sim Uniform(\overrightarrow{\Phi}_i)$, such that each node is uniformly likely to connect to every other node in a given direction.
- Edge potentials $\psi(x_j, x_i) = x_i^T W_Q^T W_K x_j$, in effect measuring the similarity of x_j and x_i' in a projected space.
- Value functions $v_i(x, \phi_i = [j]) = W_V x_j$, a linear transformation applied to the node at the start of the edge ϕ_i.

C.2. Positional Encodings and Graph Neural Networks

In Table.1 we show that positional encodings and graph attention are naturally incorporated in this framework. Absolute positional encoding as suggested by Vaswani et al. (2017) can be seen as modifying the edge potentials with a vector that depends on position, while relative position encodings can be seen as a categorical prior, where the prior depends on the relative distance between nodes. Graph and Sparse attention operate similarly to graph attention, except the uniform prior is restricted to edges in the provided graph, or according to predefined sparsity pattern.

Relative Position Encodings If the prior over edges is categorical i.e. $P(\phi_i = [j]) = p_{i,j}$, it can be fully specified by the matrix $(P)_{i,j} = p_{i,j}$. This leads to the modified attention update

$$\sum_j softmax_j(x_i Q^T K x_j + \ln p_{ij}) x_j$$

However this requires local parameters for each node z_i. A more natural prior assign a different probability to the relative distance of i from j. This is achieved with $P = circ(p_1, p_2, .., p_n)$, where $circ$ is the circulant matrix of $(p)_{i \leq n}$. Due to properties of circulant matrices $\ln P_{ij}$ can be reparameterised with the hartley transform

$$\ln p_{i,j} = \sum_k \beta_k [cos(k\theta_{i,j}) + sin(k\theta_{i,j})] = \beta \cdot b^{(i,j)}$$

Where $b_k^{(i,j)} = cos(k\frac{i-j}{2\pi n}) + sin(k\frac{i-j}{2\pi n})$ can be thought of as a relative position encoding, and β are parameters to be learnt.

C.3. Block-Slot Attention

Singh et al. (2022) suggest combining an associative memory ability with an object-centric slot-like ability and provide an iterative scheme for doing so, alternating between slot-attention and hopfield updates. Our framework permits us to flexibly combine different attention mechanisms through different latent graph structures, allowing us to derive a version of block-slot attention.

In this setting we have three sets of variables X, the observations, Z the latent variables to be inferred and M which are parameters. Define the pairwise MRF $X = \{x_1, ..., x_n\}$, $Z = \{z_1, ..., z_m\}$ and $M = \{m_1, ..., m_l\}$ with a prior over edges $p(E) = \prod_{j=1}^m p(E_j) \prod_{k=1}^l p(\tilde{E}_k)$, $E_j \sim Uniform\{(x_j, z_1), .., (x_j, z_m)\}$, $\tilde{E}_k \sim Uniform\{(z_1, m_k), .., (z_m, m_k)\}$, with edge potentials between X and Z given by $\psi(x_j, z_i) = z_i Q^T K x_j$ and between Z and M, $\psi(z_i, m_k) = z_i \cdot m_k$ applying (9) gives

$$\begin{aligned}\mu_i^* = &\sum_j softmax_i(\mu_i Q^T K x_j) Q^T K x_j \\ &+ \sum_k softmax_k(\mu_i \cdot m_k) m_k\end{aligned}$$

In the original block-slot attention each slot z_i is broken into blocks, where each block can access block-specific memories i.e. $z_i^{(b)}$ can has possible connections to memory nodes $\{m_k^{(b)}\}_{k \leq l}$. Allowing objects to be represented by slots which in turn disentangle features of each object in different blocks. We presented a single block version above, however it is easy to see that the update extends to the multiple block version applying (9) gives

$$\begin{aligned}\mu_i^* = &\sum_j softmax_i(\mu_i Q^T K x_j) Q^T K x_j \\ &+ \sum_{k,b} softmax_k(\mu_i^{(b)} \cdot m_k^{(b)}) m_k^{(b)}\end{aligned}$$

D. Experimental Details

Multihop Task Setup We simulate a simple dataset that has this property using the following data generation process: Initialise a projection matrix $W_y \in \mathbb{R}^{d \times 1}$ and a relationship matrix $W_r \in \mathbb{R}^{d \times d}$. X is then generated causally, using the relationship $x_{i+1} = W_r x_i + N(0, \sigma)$ to generate x_0, x_1 and x_2, while the remaining nodes are sampled from the noise distribution $N(0, \sigma)$. Finally, the target y is generated from the history of x_2, $y = W_y(x_1 + x_0)$ and the nodes of X are shuffled. Importantly W_r is designed to be low rank, such that performance on the task requires paying attention to both x_1 and x_0, Figure 2. Following the notation in the text; data generation parameters:

- Total number of tokens: 10
- Embedding dimension (dimension of each x):10

- Output dimension (dimension of y): 1
- σ^2 (autoregressive noise): 1
- Random matrix initialisation was performed with torch.rand

Training parameters (across all models):

- batch size:200
- number of batches: 10
- optimiser: ADAM
- learning rate: $1e-3$
- Number of different random seeds: 10

Model: To make analysis easier, all models were prevented from self-attending to the final token.

Expanding Task Setup Input and target sequence are generated similarly to above (without x_0). Here x_1 is moved away from x_2 according to a draw from a geometric distribution, Figure 3. Following the notation in the text; data generation parameters:

- Total number of tokens: 50
- Embedding dimension(s) (dimension of each x):[10, 50]
- p the parameter for generating a geometric shuffle:[0.5, .2, .1, .04]
- Output dimension (dimension of y): 1
- σ^2 (autoregressive noise): .1
- Random matrix initialisation was performed with torch.rand

Training parameters (across all models):

- batch size:1*
- number of batches: 10000
- optimiser: ADAM
- learning rate: $5e-4$

Model: To make analysis easier, all models were prevented from self-attending to the final token. For expanding attention the hyperparameters were set as $\alpha = .1$, $\beta = .9$ these were chosen to have a mean value at roughly a quarter of the (size 50) window.

*Training was performed with single samples, despite the iterative process being completely parallel (no shared state). Naive parallel implementation of expanding attention would encounter synchronisation locks, as the fastest samples wait for the longest ones to complete. In order to take full advantage of a dynamic window over a batch, intelligent asynchronous processing would be necessary.