# Attention as inference with third-order interactions

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

In neuroscience, attention has been associated operationally with enhanced processing of certain sensory inputs depending on external or internal contexts such as cueing, salience, or mental states. In machine learning, attention usually means a multiplicative mechanism whereby the weights in a weighted summation of an input vector are calculated from the input itself or some other context vector. In both scenarios, attention can be conceptualized as a gating mechanism. In this paper, we argue that three-way interactions serve as a normative way to define a gating mechanism in generative probabilistic graphical models. By going a step beyond pairwise interactions, it empowers much more computational efficiency, like a transistor expands possible digital computations. Models with three-way interactions are also easier to scale up and thus to implement biologically. As an example application, we show that a graphical model with three-way interactions provides a normative explanation for divisive normalization in macaque primary visual cortex, an operation adopted widely throughout the cortex to reduce redundancy, save energy, and improve computation.

## 1 Introduction

Attention appears to play an important role in intelligently sifting through information in a complex world. The term 'attention' is used in both neuroscience and machine learning, but the relationship between the two is largely suggestive, rather than quantitative. We propose a normative mathematical framework for modeling the natural environment that may unify these notions of attention in brains and machines.

In neuroscience, attention describes the dynamic selection of a subset of inputs for more efficient processing. It usually refers to a behavioral or perceptual phenomenon, rather than a neural mechanism, and thus has been defined operationally based on changes in performance with different kinds of cueing. We can distinguish between bottom-up and top-down attention based on the origin of the weight modulation. For instance, a visual saliency map [15, 22] is bottom-up attention since weights are calculated in a feedforward way from upstream visual signals. In contrast, covert spatial attention [28] appears to be top-down: the observer is cued by signals of a different modality or from a different time to attend more to one location than at others. Another example of covert visual attention is feature attention, where the observer is cued to enhance processing of one particular feature, like a color or shape, regardless of spatial location [26]. Importantly, experimental evidence links these perceptual effects of attention with neural mechanisms of divisive normalization [23, 25], a ubiquitous neural computation.

In machine learning, attention is a gating mechanism named to evoke the gating phenomena of natural attention. This mechanism captured substantial interest in recent years for its ability to build models that match and exceed the performance of previous recurrent models, without needing recurrence. Originally proposed in [2], models with attention nowadays like Transformers [29]can be scaled to accommodate billions of parameters, yielding superior performance on tasks including

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

machine translation [29], language generation [4, 9], and even vision tasks [10] which were previously dominated by convolutional neural networks. Attention in machine learning has a similar form as in neuroscience: a vector of original inputs are reweighted for enhanced processing with more expressive power and flexibility. Details of the attention mechanism, like the formula of weight function and the sources of weights, have evolved to maximize the training and scaling efficiency of this building block [2, 29].

Despite differences in implementation details of attentional modulation in neuroscience and machine learning, the basic motif is shared: a multiplication of the original signal by an additional controllable weight (divisive normalization is still multiplicative modulation, using a gain equal to the inverse of the normalizer). We propose that this operation arises naturally from a recurring motif in a generative model of natural inputs. Specifically, we consider probabilistic graphical models that go beyond pairwise interactions to include interactions between three variables. The simplest form of such interactions naturally yields gating-type interactions, whereby one variable controls the amount of interaction between the other two. Inference in this model generalizes past notions of flexible divisive normalization, and provides a richer expression of the normalization motif associated with attentional modulation. We argue that multiplicative third-order interactions are not only efficient building blocks for machine learning models, but also a normative way to account for the cortical mechanisms of attentional gating.

As a concrete example, in this paper we use third-order interactions to build a generalized version of a previous model that accounts for flexible divisive normalization in primary visual cortex. This earlier work argued that divisive normalization was a natural inference operation for a Gaussian Scale Mixture (GSM) [1], which is a useful component of generative models for natural scenes [30]. We will see below that our core third-order motif is already implicit in this simple model. Generalizing this approach, subsequent work [6] started to connect divisive normalization in primary visual cortex with context sensitivity, by defining a graphical model with input-dependent interactions implemented as a mixture model called a Mixture of Gaussian Scale Mixtures (MGSM) [12]. In the MGSM, the mixture weights represent whether two neurons with neighbouring receptive fields see similar patterns and if it should normalize their activities accordingly. Neural activities were linked to this normative model by the hypothesis that neuronal activities represent the posterior means inferred for the unmixed latent variables. The activity predicted by MGSM agrees with single neuron recordings when the animal is presented with controlled grating images [6, 7]. However, the discrete categories of the original MGSM were specialized for one type of problem, and even there they could not be scaled to models covering larger receptive fields without exponentially increasing complexity.

We show that by generalizing the MGSM using multiplicative third-order interactions, we can define a more interpretable and scalable model for flexible divisive normalization. When applied to low-level visual stimuli, our approach replicates the core MGSM predictions about primary visual cortex and qualitatively matches experiments [5, 16]. Furthermore, our framework can also be used to build larger models for higher-level attention models like the combination of spatial and feature attention.

# 2 Methods

## 2.1 Third-order Interaction as a Local Attention

Third-order interactions can be interpreted as a gating/reweighting operation whereby one variable controls whether and how much two others interact. Arguably, this is the essential structure of attention. We can see this structure in the GSM. A single observable image pixel $x$ is the product of the surface reflectance $y$ of an object and the illuminant $z$ along with some observation noise $\eta$. For a Gaussian noise model, we can write the observation probability in a GSM as $p(x|y,z) = \mathcal{N}(x|\mu = yz, \sigma^2) \propto \exp\left(-(x-yz)^2/2\sigma^2\right.$. The exponent expands to reveal $-\frac{1}{\sigma^2}(x^2 - 2xyz + y^2z^2)$. The key term that connects the variables is the third-order term $xyz$. Figure 1A shows a joint density contour plot highlighting this crucial term, $p(x,y,z) \propto B(x,y,z)\exp xyz$, where $B = \exp -\|(x,y,z)\|_2^4$ is an isotropic base measure. We can see that the correlation between $x$ and $y$ varies from negative to positive as we increase $z$. Figure 1B depicts the corresponding three-variable probabilistic graphical model motif with a factor capturing the third-order interaction term. Because gating between three interacting variables functions much like a transistor, we like to call this third-order motif a 'statistical transistor'. Building models with continuous variables in statistical transistors at large scale effectively constructs an attentional model.

As a general framework, we put first, second, and third-order interactions as energy terms in the exponential family [8, 20, 24]. The joint probability density of all variables $\mathbf{x}$, whether observed or not, can be written as $p(\mathbf{x}) \propto \exp(-\beta E(\mathbf{x}))$, where $E(\mathbf{x})$ is an energy function and the inverse temperature $\beta$ controls the scale of the energy landscape. The joint probability is then

$$p(\mathbf{x}) = \frac{1}{Z} \exp\Big( -\beta \Big[ \sum_i b_i x_i + \sum_{ij} a_{ij} x_i x_j + \sum_{ijk} c_{ijk} x_i x_j x_k \Big] \Big)$$

Without any constraint, the number of parameters is cubic in the number of variables, thus making learning and inference hard. In practice, we may impose a sparse task-dependent graph structure onto the model to reduce the number of parameters, like a bipartite structure for latent space model, or a grid structure when modeling translation-invariant visual processing. As an concrete example in neuroscience, we re-formulate the MGSM model for divisive normalization in primate visual cortex with our framework and get similar results, as described below.

### 2.2 Gaussian Scale Mixture and Mixture of Gaussian Scale Mixture

Gaussian Scale Mixture (GSM) is a probabilistic model describing the local variance dependency in natural images when represented in oriented basis [30]. In a vision application of a GSM, basis vectors for image patches are summed to generate a local image patch, with coefficients given by the product of a multivariate Gaussian and a shared scalar mixer variable. So that different parts of an image could have different statistical correlations, [6] employed a finite mixture of GSMs (MGSM) [12] to model relationships between neighboring image patches. Depending on the dominant component in neighboring image patches, the MGSM allows for different interactions between those patches. [6] uses a version of MGSM that explicitly distinguishes a center patch and several surrounding patches, fits it to natural image data, and uses the posterior mean of different patches to predict the neural responses. The mathematical dependence of this posterior mean takes the form of context-dependent divisive normalization [6, 7].

### 2.3 Continuous Mixture of Gaussian Scale Mixture

**Graphical Model.** The MGSM works as a divisive normalization model by being able to flexibly switch between components that represent independent or dependent pairs of center-surround image patches. While a finite mixture model is a simple way to achieve this, the number of components grows exponentially with the number of image patches if we want to build a model for the whole image instead of local patches.

In this paper we define a substantially more general model that includes a continuous MGSM as a special case. Our third-order model formulates the distribution with an energy-based model using multiplicative third-order terms as building blocks. The core motif of our graphical model is depicted in Fig. 1B, and a fuller example application replicating the center-surround geometry of [7] is shown in Fig. 1C. As in [6], there is a center GSM corresponding to the center RF and four surrounding GSMs, each corresponding to one orientation at all surrounding locations. The observable variables of each GSM are the coefficients of edge filters applied to neighboring image patches (see Fig. 1C, 'image') [27]. These filters used 4 orientations and both even and odd phases. For clarity, we only illustrate the interaction between the center GSM and the vertical surrounding GSM. Instead of using a five-state categorical variable $\xi$ to indicate which two groups of Gaussian variables $\mathbf{g}_c, \mathbf{g}_i$are co-assigned with a shared scale mixer variable $v$, here we let each group have its own mixer variable and modulate the correlation between two neighbouring mixers $v_c, v_i$ via a third-order interaction with another variable $w_i$: $\exp(-w_i v_c v_i)$. In this way, the whole graphical model is unified by third-order interactions, making it readily generalizable and more intepretable. The logarithm of unnormalized probability density (energy function) is:

$$E(\mathbf{g}, \mathbf{v}, \mathbf{w}) = \sum_i v_i^2 + \sum_{(i,j)} \alpha_{ij} w_{ij} (v_i - v_j)^2 + \sum_{(i,j)} [\mathbf{g}_i \; \mathbf{g}_j] P_{ij}(w_{ij}) [\mathbf{g}_i \; \mathbf{g}_j]^T + \sum_i N_i \log v_i \quad (1)$$

where $\boldsymbol{g}_i$ and $v_i$ are the latent Gaussian random vector and the mixer variable respectively in GSM component $i$, $w_{ij}$ is the gating variable connecting two GSMs $i$ and $j$, $P_{ij}(w_{ij})$ is the $w_{ij}$-dependent correlation matrix for the joint vector $(\boldsymbol{g}_i \; \boldsymbol{g}_j)$, and $N_i$ is the dimension of the Gaussian vector $\boldsymbol{g}_i$. The second and third sums are taken over all connected pairs $(i, j)$, which are four center-orientation pairs

in this case. The last term in Eq. (1) comes from the deterministic relation $\mathbf{k}_i = v_i \mathbf{g}_i$. Because of this, observable variables $\mathbf{k}_i$s are integrated out with the delta functions and thus not shown in Eq. (1). Some additional quadratic terms are added to the third-order model to ensure that the coupling will approximately preserve the marginal variance of each mixer variable. For a center-surround pair $(\mathbf{g}_c, \mathbf{g}_i)$ where $c$ stands for center and $i$ represents one of the four surrounding GSMs, we let the joint "precision matrix" $P_{ci}(\{w_{ci}\})$ (see Fig. 1C) be linearly dependent on $w_{ci}$, and normalized by all local $w_{cj}$'s,so that it ranges from an independent block-diagonal structure to a dense positive-definite matrix as $w_{ci}$ goes from zero to infinity.

**Learning.** For our model, we are able to calculate the normalization constant as a function of all parameters, up to a constant. So even if the constant is unknown, we can still use Evidence Lower Bound (ELBO) as our loss function for variational inference [19]. A simple variational encoder [19] is used as the variational posterior for the ELBO training which maps the input/observable vector $\mathbf{k}$ into a mean vector along with a variance vector and models the posterior distribution as independent Gaussian variables given the encoded mean and variance. The model is trained with stochastic gradient descent using the negative ELBO as objective, evaluated on 500 natural images from the Berkeley Segmentation Dataset (BSDS500). To construct the training dataset, we randomly sample 100,000 image patches from BSDS500 and then pass them to the Steerable Pyramid to generate the inputs for the graphical model, $\mathcal{D} = \{\mathbf{k}\}_{1..100,000}$. We use the Adam optimizer [18] with initial learning rate $0.001$ and batch size 500.

**Inference.** It is not easy to infer the latent variables in our third-order model. One option is to train Graph Networks to provide approximate inferences [11, 31]. However, for accuracy, here we infer the latent variables conditioned on each stimulus by using a Markov Chain Monte Carlo sampling algorithm. Specifically, we use the 'No U-Turn Sampler' NUTS [13] implementation from the Pyro package [3].

**Link to neural data.** We follow [6] in hypothesizing that neural activity $r$ reflects the marginal posterior means of the underlying latent Gaussian variables $\boldsymbol{g}_i$. More precisely, for each orientation in the center RF, $\mathbf{k}$ and $\mathbf{g}$ each have two components that correspond to even and odd phase. We combine the posterior means of these two Gaussian variables to achieve a phase invariant prediction consistent with complex cells in primary visual cortex [14]:

$$r_{\mathrm{c,vertical}} = \sqrt{(\mathbb{E}_{p(\mathbf{g}_c|\mathbf{k})} g_{\mathrm{c,vertical,even}})^2 + (\mathbb{E}_{p(\mathbf{g}_c|\mathbf{k})} g_{\mathrm{c,vertical,odd}})^2} \tag{2}$$

# 3 Results

After our model is trained, we test if our model built with third-order interactions behaves like the MGSM in predicting experimentally measured neural activity. We performed two grating experiments originally used in [5, 16] and employed in [6] for comparison.

For the first experiment, the input image consists of a neutral grey background with a circular patch of a vertical grating in the middle.The diameter of the center patch is increased during the experiment. For the second experiment, the input image again has a center patch with a vertical grating covering the entire center RF, large enough to partly overlap the surrounding RFs. Outside of the center patch is an annulus covering most of the surrounding RFs, which is filled by a grating pattern at another orientation. The orientation of the surrounding grating is varied during the experiment, and for each orientation we average over five spatial phases, effectively changing the correlation between the center stimulus and the surrounding stimuli. Small visual illustrations of the input stimuli are displayed along the axes in Fig. 1C.

The neural activity predicted by our inference model (Section 2.3) exhibits similar contextual modulation and divisive normalization effects as recorded in V1 neurons, as shown in Fig. 1D,E. For compactness we omit the predictions by the original MGSM that motivated these comparisons and which our results recapitulate, but these plots show the same behavior as can be seen in Figure 7 and 8A of [6]. The agreement between neural data and our model provides support for the idea that the third-order motif may be widely applicable in neuroscience, and can provide a normative account of a more generalized form of flexible, context-dependent divisive normalization and attention.

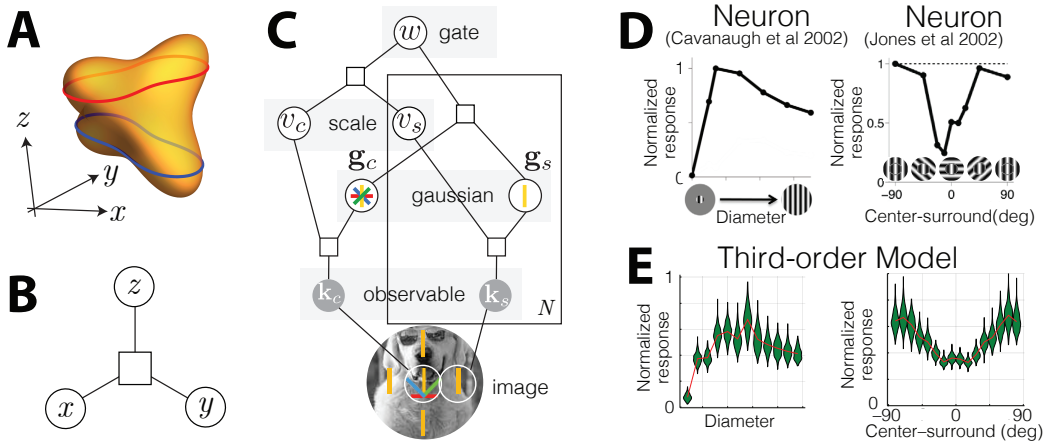


Figure 1: Third order interactions. **A**: Contour of the third-order joint distribution $p(x, y, z) \propto be^{xyz}$, with isotropic base measure $b(x, y, z) = \exp\left[-\|(x, y, z)\|_{\ell_2}^4\right]$. One variable can be viewed as modulating the interaction between the other two. For example, $x$ and $y$ are negatively correlated for small values of $z$ (blue contour), but are positively correlated when $z$ is large (red contour). **B**: Third-order motif as graphical model. **C**: Graphical model between the center Receptive Field (RF) and one surrounding RF in a third-order version of an MGSM. The large rectangle denotes a plate, indicating multiple copies of the enclosed variables. **D**: Neural tuning in two grating experiments. *Left*: Neuron with vertical tuning in center RF responds with the same vertical patterns when presented with round vertical stimuli with increasing diameter. *Right*: For a stimulus with a center grating covering the whole center RF and partially covering the surround RF, the normalization depends on the angle between center and surround stimuli. **E**: Similar trends are predicted by our model (as for the MGSM [6], not shown).

## 4 Discussion

Our third-order motif substantially generalizes the MGSM, showing that third-order interactions offer a more flexible way to describe context-dependent divisive normalization, a modulatory effect that functions as a form of attention [23, 25]. In addition, we can also construct graphical models for other types of attention. For example, while our third-order model only accounts for attention locally, it could be easily expanded to a larger scale by forming a grid of variables connected by third-order interactions. Performing inference on this expanded model may produce spatial attention, and has the potential to provide interpretable neural computations involved in tasks like object segmentation [12, 21].

We can also construct a hierarchical graphical model by stacking layers together with third-order interactions in between as a soft local gating/attention to model top-down attention or bottom-up attention. Attention is often directed at task-relevant variables, and determining the task relevance can be viewed as an inference problem. In particular, we can appeal to the control-as-inference framework [17], where the task of maximizing subjective value is transformed into the task of inferring which actions are most probable given that a latent indicator of optimality is true. Task cues could then influence these latent indicators through the hierarchical graphical model structure, and could generate inferences equivalent to spatial or feature-based attention.

Attention in neuroscience and machine learning is closely connected to gating and divisive normalization. We showed that these ideas can emerge as natural inferential computations in third-order models. Here we do not suggest a mechanism for how this computation is carried in the brain, but instead offer a normative model of such computation, assuming that the brain is performing Bayesian inference using third-order statistical relationships. However, the information flow in algorithms to implement this Bayesian inference may serve as hypotheses for mechanisms of distributed attention-related computation [11], and is an important topic for future investigation.

# References


[1] David F Andrews and Colin L Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):99–102, 1974.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] James R Cavanaugh, Wyeth Bair, and J Anthony Movshon. Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. *Journal of neurophysiology*, 88(5):2530–2546, 2002.

[6] Ruben Coen-Cagli, Peter Dayan, and Odelia Schwartz. Cortical surround interactions and perceptual salience via natural scene statistics. *PLoS Comput Biol*, 8(3):e1002405, 2012.

[7] Ruben Coen-Cagli, Adam Kohn, and Odelia Schwartz. Flexible gating of contextual influences in natural vision. *Nature neuroscience*, 18(11):1648–1655, 2015.

[8] Georges Darmois. Sur les lois de probabilitéa estimation exhaustive. *CR Acad. Sci. Paris*, 260(1265):85, 1935.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Yicheng Fei and Xaq Pitkow. Generalization of graph network inferences in higher-order probabilistic graphical models. *arXiv preprint arXiv:2107.05729*, 2021.

[12] Jose A Guerrero-Colón, Eero P Simoncelli, and Javier Portilla. Image denoising using mixtures of gaussian scale mixtures. In *2008 15th IEEE International Conference on Image Processing*, pages 565–568. IEEE, 2008.

[13] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[14] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

[15] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[16] HE Jones, W Wang, and AM Sillito. Spatial organization and magnitude of orientation contrast interactions in primate v1. *Journal of neurophysiology*, 88(5):2796–2808, 2002.

[17] Hilbert J Kappen, Vicenç Gómez, and Manfred Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[20] Bernard Osgood Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3):399–409, 1936.

[21] Miguel Lázaro-Gredilla, Wolfgang Lehrach, Nishad Gothoskar, Guangyao Zhou, Antoine Dedieu, and Dileep George. Query training: Learning a worse model to infer better marginals in undirected graphical models with hidden variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8252–8260, 2021.

[22] Zhaoping Li. A saliency map in primary visual cortex. *Trends in cognitive sciences*, 6(1):9–16, 2002.

[23] Amy M Ni, Supratim Ray, and John HR Maunsell. Tuned normalization explains the size of attention modulations. *Neuron*, 73(4):803–813, 2012.

[24] Edwin James George Pitman. Sufficient statistics and intrinsic accuracy. In *Mathematical Proceedings of the cambridge Philosophical society*, volume 32, pages 567–579. Cambridge University Press, 1936.

[25] John H Reynolds and David J Heeger. The normalization model of attention. *Neuron*, 61(2):168–185, 2009.

[26] Andrew F Rossi and Michael A Paradiso. Feature-specific effects of selective visual attention. *Vision research*, 35(5):621–634, 1995.

[27] Eero P Simoncelli and William T Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proceedings., International Conference on Image Processing*, volume 3, pages 444–447. IEEE, 1995.

[28] Roger BH Tootell, Nouchine Hadjikhani, E Kevin Hall, Sean Marrett, Wim Vanduffel, J Thomas Vaughan, and Anders M Dale. The retinotopy of visual spatial attention. *Neuron*, 21(6):1409–1422, 1998.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[30] Martin J Wainwright and Eero Simoncelli. Scale mixtures of gaussians and the statistics of natural images. *Advances in neural information processing systems*, 12, 1999.

[31] KiJung Yoon, Renjie Liao, Yuwen Xiong, Lisa Zhang, Ethan Fetaya, Raquel Urtasun, Richard Zemel, and Xaq Pitkow. Inference in probabilistic graphical models by graph neural networks. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 868–875. IEEE, 2019.