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. Generally, attention has been conceptualized as a spotlight [26] that gates or reweights aspects of the incoming signal using a controllable weight. 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 [30] 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 [28]. Importantly, experimental evidence links these perceptual effects of attention with neural mechanisms of divisive normalization [24, 27], 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.

We show that by generalizing the MGSM using multiplicative third-order interactions, we can define a model motif with a factor capturing the third-order interaction term. Because gating between three variables functions much like a transistor, we like to call this third-order motif a ‘statistical model motif with a factor capturing the third-order interaction term.

As a concrete example, in this paper we use third-order interactions to build a generalized version 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 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 [32]. 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 illuminance $z$ along with some observation noise $\eta$. For a Gaussian noise model, we can write the observation probability in a GSM as $p(x|y, z) = N(x|\mu = yz, \sigma^2) \propto \exp(-\frac{1}{2}(x - yz)^2/2\sigma^2)$. The exponent expands to reveal $-\frac{1}{2}(x^2 - 2xyz + y^2z^2)$. 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, $p(x, y, z) \propto B(x, y, z) \exp x y z$, where $B = \exp -\| (x, y, z) \|^2$ 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, 25]. The joint probability density of all variables \( x \), whether observed or not, can be written as \( p(x) \propto \exp(-\beta E(x)) \), where \( E(x) \) is an energy function and the inverse temperature \( \beta \) controls the scale of the energy landscape. The joint probability is then

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

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 [32]. In a vision application of a GSM, oriented 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 orientation groups in surrounding patches, fits it to natural image data, and uses the posterior mean of different orientations in the center patch to predict the neural responses. The mathematical dependence of this posterior mean takes a 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 in 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’) [29]. 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 \( g_c, g_v \) 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 \( w_{ij} \): \( \exp(-w_{ij} v_c v_i) \). In this way, the whole graphical model is unified by third-order interactions, making it readily generalizable and more interpretable. The logarithm of unnormalized probability density (energy function) is:

\[
E(g, v, w) = \sum_i v_i^2 + \sum_{i,j} \alpha_{ij} w_{ij} (v_i - v_j)^2 + \sum_{i,j} [g_i, g_j] P_{ij}(w_{ij}) [g_i, g_j]^T + \sum_i N_i \log v_i \tag{1}
\]

where \( g_i \) and \( v_i \) are the unmixed 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...
which our results recapitulate, but these plots show the same behavior as can be seen in Figure 7 and
with initial learning rate 0.001. The neural activity predicted by our inference model (Section 2.3) exhibits similar contextual
variables conditioned on each stimulus by using a Markov Chain Monte Carlo sampler, specifically
variational encoder [19] is used as the variational posterior for the ELBO training. This encoder
is an annulus covering most of the surrounding RFs, which is filled by a grating pattern at another
originally used in [5, 16] and employed in [6] for comparison. MGSM in predicting experimentally measured neural activity. We performed two grating experiments
in GSM. Because of this, observable variables $k_{i,s}$ 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 $(g_c, g_i)$ where $c$ stands for center and $i$ represents one of the four surrounding
GSMs, we let the joint “precision matrix” $P_{\text{ci}}(\{w_{ci}\})$ (see Fig. 1C) be linearly dependent on $w_{ci}$, and normalized by all local $w_{ci}$’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. This means that even if the constant is unknown, we can still use
the 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. This encoder
maps the input/observable vector $k$ into a mean vector and scale vector, and models the posterior
distribution as a product of independent distributions given the encoded mean and scale. Here we
assume these component distributions are log-normal to enforce a constraint that all latent variables
are non-negative; future work will explore relaxing this constraint. The model is trained with
stochastic gradient ascent using the ELBO objective evaluated on 500 natural images from the
Berkeley Segmentation Dataset (BSDS500)[23]. 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, $D = \{k\}$...100,000. We use the Adam optimizer [18]
with initial learning rate 0.001 and batch size 2000.

**Inference.** It is hard to infer the latent variables in our third-order model. One option is to train Graph
Networks to provide approximate inferences [11, 33]. However, for accuracy, here we infer the latent
variables conditioned on each stimulus by using a Markov Chain Monte Carlo sampler, specifically
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 $\mathbf{g}_i$. More precisely, for each orientation
in the center RF, $\mathbf{k}$ and $\mathbf{g}$ each have two components that correspond to even and odd phases. We
combine the posterior means of these two Gaussian variables to achieve a phase-invariant predicted
response, consistent with complex cells in primary visual cortex [14]:

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

(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
Figure 1: Third-order interactions. A: Contour of the third-order joint distribution $p(x, y, z) \propto b e^{xyz}$, with isotropic base measure $b(x, y, z) = \exp \left[-\|x, y, z\|_{L^2}^2\right]$. 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).

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.

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 [24, 27]. 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.
Acknowledgements  This work was supported by NSF NeuroNex grant 1707400 to XP.

Conflicts of interest.  XP is a founding member of Upload AI, LLC.

References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 2.1
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Training code and grating stimuli data could be provided upon request
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 2.3
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report the full posterior distribution of inferred Gaussian variables as violin plots in Fig. 1E
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The training takes less than an hour on an Apple Macbook Pro with M1 Pro chip

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use the BSDS500 dataset [23] and the code provided by [6]
   (b) Did you mention the license of the assets? [No] BSDS is free for noncommercial research purposes
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]