Attention as reward-driven optimization of sensory processing

Matthew Chalk, Iain Murray & Peggy Seriès

1 Summary of model assumptions

In order to make concrete predictions about how attention should alter visual neuron responses, we made had to make certain assumptions about the neural code, agent's internal model, and the modulatory effects of attention. Here we summarize the assumptions that are critical for our results, providing a brief theoretical justification, and outlining how each assumptions influences the results of our simulations.

- **Neural code.** We assume a very simple neural code, in which the firing rate of each visual neuron is directly proportional to the posterior probability that a single latent variable is active. This choice of code is important for our simulations, as it leads to an expression for neural firing rates which has a similar functional form to previous divisive normalization models of neural responses [1, 2]. In these models, neural firing rates are evaluated by dividing their feedforward excitatory drive by a suppressive factor that depends on the summed activity nearby neurons. In our model, divisive normalization emerges from Bayes' law, due to the fact that the posterior probability for a hidden variable to be active ($p(y_i = 1|\boldsymbol{x})$) is evaluated by dividing the joint distribution over hidden and observed variables ($p(y_i = 1, \boldsymbol{x})$) by the marginal probability for the observed sensory input ($p(\boldsymbol{x})$).

- **Sparse stimulus statistics.** We assume that the agent learns a 'sparse' internal model, in which there is a small prior probability for any particular hidden cause to be active (i.e. $p(y_i = 1|\theta) \ll 1$). The theoretical justification for this comes from natural image statistics, which are well accounted for by sparse models [3, 4]. In our simulations, the sparse prior produces strong competition between different explanations of the received sensory input, which is reflected in the divisive suppression of neural responses.

- **Non-linear combination rule.** The agent learns an internal model in which stimuli are assumed to combine non-linearly, according to a 'max' combination rule. While many previous generative models of visual processing have assumed a linear combination rule, arguably, a nonlinear 'max' combination rule provides a better description of how features combine in natural images [5]. In our simulations, a nonlinear combination rule was required to produce neural responses that saturated below their maximum values when the stimulus contrast was high (see figure 7). In contrast, a linear rule results in contrast response plots that go towards 1 when the strength of the sensory input is increased (supplementary figure 1).

  To see why this is the case consider what happens when the agent receives a high amplitude sensory input (i.e. components of x have high values). A high amplitude

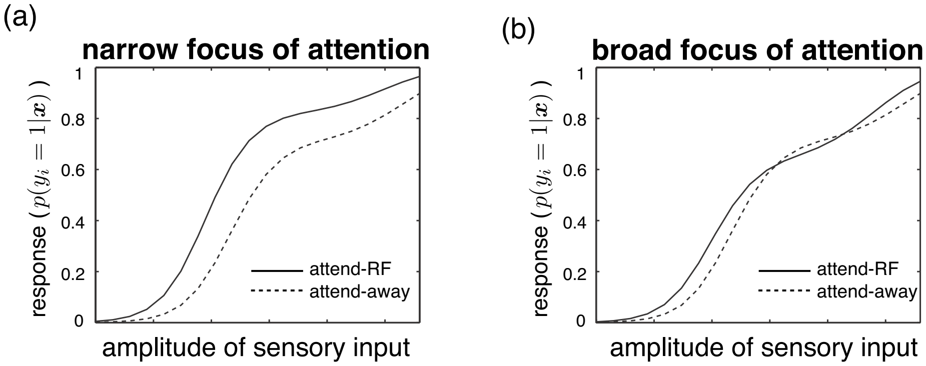

Figure 1: Contrast response plots, obtained using a model in which stimuli combine linearly to activate the sensory inputs. All other simulation details are identical to figure 7 in the main text.

sensory input can be well explained by a linear model (where $x_i = \sum_j A_{ij} y_j + \gamma_i$) if multiple hidden units are active. Thus, if the agent uses a linear internal model, they will ascribe a high posterior probability for multiple hidden to be active simultaneously (i.e. the posterior probability, $p\left(y_i = 1|\boldsymbol{x}\right)$, will saturates at ~ 1). However, with a nonlinear 'max' rule ($x_i = \max_j\{A_{ij}^{true} y_j\} + \gamma_i$), multiple hidden variables do not combine to produce a higher amplitude sensory input. Thus, different hidden variables compete to explain the data. Thus, if the agent uses an internal model with a 'max' combination rule, the posterior probability that they ascribe to individual hidden units will remain below 1, even when the amplitude of the sensory input is high.

- **Internal model structure.** The agent learns a hierarchical internal model, in which high-level hidden variables ($\boldsymbol{z}$), that correspond to the global structure of the sensory input, are assumed to determine the state of low-level hidden ($\boldsymbol{y}$), that correspond to the local features of the sensory input. This model structure is designed to reflect the structure of natural images, in which complex objects are made up of simple image features, which give rise to the observed sensory input [6, 7]. The agent assumes that only the high-level latent variables determine the reward that will be received for performing different actions. This model structure could allow the agent to quickly adapt to new behavioural contexts, as the action that they should perform in any given task will depend on a limited number of high-level hidden variables. However, in our simulations, we investigate a situation when this internal model structure is suboptimal: when the image features that are relevant to the task are more spatially localized than the high-level features in the agent's internal model. In our work, it is this mismatch between the structure of the agent's internal model and the behavioural task that drives attentional modulation of visual neuron responses.

  Note that while we assumed a small number of high-level variables ($n_z = 5$), our qualitative results are not highly sensitive to the number of z-units. For example, supplementary figure 2 shows that qualitatively similar are obtained with twice the number of z-units ($n_z = 10$). However, in some cases, increasing the number of z-units will lead to slower learning of the task, as the agent must learn a mapping

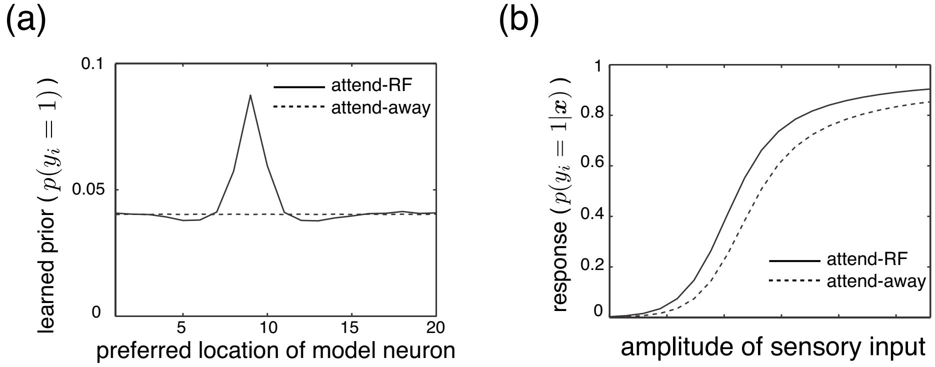

Figure 2: Results obtained with 10 z-units (all simulations in the main text have only 5 z-units). (a) Prior probability assumed by the agent that each of the hidden causes are active, with or without attention. (b) Model neuron response, versus the amplitude of a sensory input centred on the neuron's preferred location.

between a larger number of high-level variables, and the reward that they receive for performing an action.

- **Attention only alters 'bias-terms' in the internal model.** We postulate that over the short timescales associated with visual attention, the prior probability that individual hidden variables are active can vary (determined by the bias terms, $\boldsymbol{b}_0$; although see previous section), while the image features represented by the latent variables (determined by the basis functions) is fixed. As a result, attention modulates the responses of model neurons to presented stimuli, but does not fundamentally change their stimulus selectivity. Our assumption can be justified functionally from the fact that updating the bias terms only requires estimating first-order statistics, which can be evaluated quicker and more reliably than the second-order statistics that are required to update the basis functions. More generally, how much attention should alter different aspects of the agent's internal model will depend on several different factors, including the rate at which different aspects of the world vary, and the trade-off between short-term optimization in a specific task, and generalization across many different tasks.

- **Binary latent variable model.** We modeled the stimulus statistics using a binary latent variable model. We chose this form of model for simplicity, in order to simulate a simple task where where subjects were rewarded for correctly detecting a stimulus. Recent work suggests that a binary latent variable model also provides a good description of natural scene statistics [8]. However, it is worth considering how this choice of model could have affected our simulations of neural responses as a function of varying stimulus contrast. In figure 6 we plot the responses of model neurons while continuously varying the amplitude of the sensory input. As stimulus contrast is not represented by the agent, increasing the amplitude of the sensory input is interpreted as increased evidence that a binary latent variable is 'on', and the model neuron response (given by $r_i \propto p\left(y_i = 1|\boldsymbol{x}\right)$) increases towards a saturating value. In a more sophisticated model, the agent could learn a joint distribution describing both the probability that a stimulus is present and its contrast. Now, if we assume that the primary goal of the early visual system is to encode the components that make up an image, we should integrate over all possible stimulus contrasts to recover the distribution $p\left(y_i = 1|\boldsymbol{x}\right)$. Intuitively, such a model should produce qualitatively similar results - a high amplitude sensory input will still indicate a high probability that a stimulus is present. However, future theoretical work will be required to see whether this is true, and thus whether our results generalize to an internal model that includes contrast as a latent variable.

2 Gradient of the objective function

To update the parameters of the agent's internal model, we need to compute the gradient of the online objective function:

$$\partial l\left(\theta, \psi\right) = \partial \log p\left(r|a, x, \theta, \psi\right).$$

The derivative of the objective function can be written as:

$$\begin{aligned} \partial l(\theta, \psi) &= \frac{1}{p\left(r|a,x,\theta,\psi\right)} \partial p\left(r|a,x,\theta,\psi\right) \\ &= \frac{1}{p\left(r|a,x,\theta,\psi\right)} \partial \int p\left(s, r|a,x,\theta,\psi\right) ds. \end{aligned}$$

Taking the derivative inside the integral,

$$\begin{aligned} \partial l\left(\theta, \psi\right) &= \frac{1}{p\left(r|a,x,\theta,\psi\right)} \int \partial p\left(s,r|a,x,\psi,\theta\right) ds \\ &= \frac{1}{p\left(r|a,x,\theta,\psi\right)} \int p\left(s,r|a,x,\psi,\theta\right) \partial \log p\left(s,r|a,x,\psi,\theta\right) ds \\ &= \int p\left(s|r,a,x,\theta,\psi\right) \partial \log p\left(s,r|a,x,\psi,\theta\right) ds, \end{aligned}$$

where we have used the identity, $\partial f\left(x\right) = f\left(x\right) \partial \log f\left(x\right)$. Rearranging this expression gives,

$$\begin{aligned} \partial l\left(\theta, \psi\right) &= \int p\left(s|r,a,x,\theta,\psi\right) \partial \log p\left(r,s|a,x,\psi,\theta\right) ds \\ &= \left\langle \partial \log p\left(r,s|a,x,\psi,\theta\right) \right\rangle_{p(s|x,r,a,\theta,\psi)} \\ &= \left\langle \partial \log p\left(r|s,a,x,\psi,\theta\right) \right\rangle_{p(s|x,r,a,\theta,\psi)} + \left\langle \partial \log p\left(s|x,\theta\right) \right\rangle_{p(s|x,r,a,\theta,\psi)}. \end{aligned}$$

The second term in this expression can be expanded as:

$$\begin{aligned} \left\langle \partial \log p\left(s|x,\theta\right) \right\rangle_{p(s|x,r,a,\theta,\psi)} &= \left\langle \partial \log p\left(s,x|\theta\right) \right\rangle_{p(s|x,r,a,\theta,\psi)} - \partial \log p\left(x|\theta\right) \\ &= \left\langle \partial \log p\left(s,x|a,\theta\right) \right\rangle_{p(s|x,r,a,\theta,\psi)} - \left\langle \partial \log p\left(s,x|\theta\right) \right\rangle_{p(s|x,\theta)}, \end{aligned}$$

where we have used the identity, $\partial \log p\left(x|\theta\right) = \left\langle \partial \log p\left(s,x|\theta\right) \right\rangle_{p(s|x,\theta)}$. Substituting this back into the expression for the derivative of the objective function, gives:

$$\begin{aligned} \partial l\left(\theta, \psi\right) = {} & \left\langle \partial \log p\left(r|s,a,\psi\right) \right\rangle_{p(s|x,r,a,\theta,\psi)} \\ & + \left\langle \partial \log p\left(s,x|a,\theta\right) \right\rangle_{p(s|x,r,a,\theta,\psi)} - \left\langle \partial \log p\left(s,x|\theta\right) \right\rangle_{p(s|x,\theta)}. \end{aligned}$$

Finally, taking the partial derivative with respect to either θ or ψ returns the expressions shown in the main text:

$$\partial_\psi l(\theta, \psi) = \left\langle \partial_\psi \log p\left(r|a,s,\psi\right) \right\rangle_{p(s|x,r,a,\theta,\psi)} \tag{1}$$

$$\partial_\theta l(\theta, \psi) = \left\langle \partial_\theta \log p\left(s,x|\theta\right) \right\rangle_{p(s|x,r,a,\theta,\psi)} - \left\langle \partial_\theta \log p\left(s,x|\theta\right) \right\rangle_{p(s|x,\theta)}. \tag{2}$$

References

[1] Reynolds, J. H. & Heeger, D. J. The normalization model of attention. *Neuron* **61**, 168–185 (2009).

[2] Carandini, M., Heeger, D. J. & Movshon, J. A. Linearity and normalization in simple cells of the macaque primary visual cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **17**, 8621–8644 (1997).

[3] Berkes, P., Turner, R. & Sahani, M. On sparsity and overcompleteness in image models. *Advances in Neural Information Processing Systems* **21** (2007).

[4] Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).

[5] Lücke, J. & Sahani, M. Maximal causes for non-linear component extraction. *The Journal of Machine Learning Research* **9**, 1227–1267 (2008).

[6] Karklin, Y. & Lewicki, M. S. A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Computation* **17**, 397–423 (2005).

[7] Reichert, D. P., Series, P. & Storkey, A. A hierarchical generative model of recurrent object-based attention in the visual cortex. In *ICANN*, 18–25 (2011).

[8] Puertas, G., Bornschein, J. & Lücke, J. The maximal causes of natural scenes are edge filters. In *Advances in Neural Information Processing 23*, 1939–1947 (2010).