Optimal Coding Predicts Attentional Modulation of Activity in Neural Systems

Santiago Jaramillo[*†] **Barak A. Pearlmutter**[*‡]

Abstract

Neuronal activity in response to a fixed stimulus has been shown to change as a function of attentional state, implying that the "neural code" also changes with attention. We propose an information-theoretic account of such modulation, namely that the nervous system adapts to optimally encode sensory stimuli while taking into account the changing relevance of different features. We show using computer simulation that such modulation emerges in a coding system informed about the uneven relevance of the input features. We present a simple feedforward model that learns a covert attention mechanism, given input patterns and coding fidelity requirements. After optimization, the system gains the ability to reorganize its computational resources (and coding strategy) depending on the incoming attentional signal, without the need of multiplicative interaction or explicit gating mechanisms between units. The modulation of activity for different attentional states matches that observed in a variety of selective attention experiments. This model predicts that the shape of the attentional modulation function can be strongly stimulus-dependent. The general principle presented here accounts for attentional modulation of neural activity without relying on special-purpose architectural mechanisms dedicated to attention. This principle applies to different attentional goals, and its implications are relevant for all modalities in which attentional phenomena are observed.

1 Introduction

Adaptation in the nervous system occurs at several time scales, from the fast changes in spiking patterns to the long-term effects of development and learning. One of the most intriguing types of adaptation is the modification of neural coding strategies related to selective attention. The fact that attention can be covertly shifted to different aspects of a stimulus to improve their detection and discrimination (Downing, 1988) indicates that the nervous system is not just filtering out features in a fixed manner, but adapting its function according to some objective. This adaptation is expressed as an attentional-state dependent modulation of neural activity. Attentional modulation of activity has been characterized at different levels, from whole brain (Heinze et al., 1994; Hopfinger et al., 2000; Corbetta and Shulman, 2002) to single cell (Moran and Desimone, 1985; Luck et al., 1997; Treue and Maunsell, 1999; McAdams and Maunsell, 1999). In contrast, theoretical accounts of these phenomena are still a matter of debate, and would greatly benefit from models that relate these effects to general coding principles.

Traditional accounts of receptive field formation assume equal relevance of all features of a stimulus. Attempts to include the adaptability necessary for dealing with uneven relevance (*e.g.*, for attention-related phenomena) usually include mechanisms tailored to the precise phenomenology, such as gating or shifting circuitry (Olshausen et al., 1993; Deco and Zihl, 2001; Heinke and Humphreys, 2003). A different approach is to find high-level principles that give rise to the phenomena observed during changing attentional state. The literature provides a few examples of this approach, namely models in which

[*] Hamilton Institute, NUI Maynooth, Co. Kildare, Ireland.

[†] sjara@ieee.org. Present address: Cold Spring Harbor Laboratory, NY 11724, USA.

[‡] barak@cs.nuim.ie

various types of uncertainty are included in a framework of optimal inference and learning (Dayan and Zemel, 1999; Dayan et al., 2000; Yu and Dayan, 2005; Rao, 2005). Our proposal here is in some sense less radical than the assumption that attention subserves optimal inference, in that we account for the sensory codes in question using the optimal coding framework, a framework which has been successfully applied to a variety of non-attentional sensory coding phenomena (Atick, 1992).

We hypothesize that the nervous system uses an optimal code for representing sensory signals, and that fidelity requirements of the code change based on top-down information. These conditions imply shifts in the neural code and changes in response properties of single neurons that altogether can be regarded as a global resource allocation. Using computer simulation we explore the efficient representation of sensory input under the assumption of limited capacity and changing nonuniform fidelity requirements.

These simulations show that: **(i)** The modulation of activity of the simulated units matches that observed in animals during selective attention tasks. **(ii)** The magnitude of this modulation depends on the capacity of the neural circuit. **(iii)** The behavior of a single neuron cannot be well characterized by measurements of attentional modulation of only a single sensory stimulus. **(iv)** Modulation of coding strategies does not *require* special mechanisms: it is possible to obtain dramatic modulation even when signals informing the system about fidelity requirements enter the system in a fashion indistinguishable from sensory signals. **(v)** Even a simple feed-forward network can perform sophisticated and dramatic reallocation of processing resources.

2 Methods

Network structure. An auto-associative network was constructed consisting of five layers connected in a feedforward fashion (Fig. 1B), following Jaramillo and Pearlmutter (2004). The number of units in each layer was 256–20–10–20–256, respectively. Each unit received inputs from all units in the previous layer, in addition to two attentional signals, displayed as a single arrow per layer in Fig. 1B. This attentional input was the same for all layers and its role depended on the cost function to be minimized as explained below. There were no lateral connections within units in a layer. The activity in the input and output layers was represented as an image of 16×16 pixels.

Unit model. A firing rate model was used in which the output of each unit was calculated as the weighted sum of the inputs passed through a saturating nonlinearity, as follows:

$$\begin{aligned} r_i &= s\Big(\sum_j w_{ij}\, r_j + b_i\Big) \\ &= s\Big(\sum_{j'} w_{ij'}\, r_{j'} + \sum_{j''} w_{ij''}\, r_{j''} + b_i\Big) \end{aligned} \tag{1}$$

where the sum is over all units from the previous layer (indexed by j') and the attentional inputs (indexed by j''). The saturating function was $s(x) = a\tanh(bx)$ with $a = 1.716$, $b = 0.667$. The activity of unit i is denoted r_i and the parameters w_{ij} correspond to the strength of the connection from unit j to unit i. Note that each unit i also included a bias term b_i. The connection strength w_{ij} from unit j to unit i was real valued and unbounded.

Stimulus statistics. The set of patterns used during optimization consisted of $20,000$ monochromatic $16{\times}16$ pixel images. Pixel intensity values had zero mean and standard deviation $\sigma = 1/3$. The images were created by convolving (filtering) white Gaussian noise images with a rotationally symmetric 2D Gaussian with $\sigma_{\mathit{filter}} = 2$. Edge effects were avoided by extracting only the 16×16 center of the resulting image. These images were then scaled to have the desired variance.

Attentional input. The attentional signal consisted of a two-element vector with elements in the range $[-1, 1]$. For each optimization step, this input was randomly drawn from a uniform distribution over the possible range. After optimization, this signal enters each unit in the same fashion as signals from other layers. It is only through the optimization process that its role is defined.

Optimization. The optimization process consisted of finding the set of weights and bias parameters that minimize the cost function $E = \langle E_{\mathbf{p}} \rangle$ where the error for one input pattern and attentional state is

$$E_{\mathbf{p}} = \sum_k c_k(\mathbf{p}) \left(y_k(\mathbf{p}) - d_k(\mathbf{p}) \right)^2 \qquad (2)$$

in which k indexes locations in the 16×16 grids holding the stimulus and its reconstruction, $c_k(\mathbf{p})$ defines the importance of that particular location (analogous to the intensity of an attentional spotlight), $y_k(\mathbf{p})$ is the output of the network, $d_k(\mathbf{p})$ is the desired output, which is in our case the same as the input, and $\mathbf{p}$ represents the complete information coming into the system at one point in time, *i.e.* the input image as well as the top-down attentional signal. The expectation $\langle \cdot \rangle$ is taken over $\mathbf{p}$.

The gradient was calculated using backpropagation (Rumelhart et al., 1986) of the weighted error defined in Equation 2, and optimization used online gradient descent with a weight decay term of 10^{-6} and a learning rate $\eta = 0.005$. All weights were plastic during learning, and the attention coefficients in the penalty function formed a simple soft mask:

$$c_k(\mathbf{p}) = \frac{1}{1 + m^2 \, \|\vec{k} - \vec{a}(\mathbf{p})\|^2} \qquad (3)$$

with $\vec{a}(\mathbf{p})$ being the attentional input (in our case, a two-dimensional vector representing the center of attention) and $\vec{k}$ being a location in the plane. The width of the attentional spotlight was set by m, which was held constant at $m = 12$ in our simulations.

Noise was injected into the bottleneck units (before the nonlinearity) during optimization in order to limit the capacity of the system. The noise level was $\sigma_{\mathrm{NB}} = 0.1$, except in Fig. 7C and D, in which additional values in the range $\sigma_{\mathrm{NB}} \in [0.0125, 1.6]$ were also used.

Testing performance. The performance of the system at encoding and decoding the input was measured using independently generated random patterns. For performance comparison (Fig. 1C), a network that used a flat penalty function $c_k(\mathbf{p}) = 1$ was also trained. The error in Fig. 1 and 7 was the absolute difference between input and output pixel values.

Finding preferred stimuli. To calculate the stimuli that maximally drive each unit in the bottleneck layer for a particular attentional state, we first generated 10^6 random white Gaussian images and found the activation of the unit of interest for each of these patterns (keeping the attentional state fixed). The preferred stimulus was defined as the average of these random patterns weighted by the activity produced in the unit of interest. The anti-preferred stimulus was defined as the negative of the preferred stimulus.

3 Results

The task of the network of Fig. 1B was to encode its input (a monochromatic image) into a compressed representation, and decode that representation, with minimal error. An additional input (a two-element vector denoted in Fig. 1B as **A**) informed the system about the current attentional state. Attentional states differed in that selected regions were preferentially reconstructed with higher quality than the rest, but these regions were not explicitly represented by any signal after optimization. Each unit computed its output as the weighted sum of its inputs passed through a saturating nonlinearity. The results presented in the following sections correspond to measurements on the system after the optimization procedure has found the appropriate connection strengths. Any modulation is due only to changes in activity and not to changes in the structure or connection strengths of the network.

Reallocation of resources emerged naturally

First, it was verified that some regions of the input image were in fact reconstructed better than others depending on the attentional state. An example of this uneven performance is presented in Fig. 1A. Here, the input image remained fixed as the attentional input changed. The dashed circles indicate those regions for which preferential reconstruction was requested. We should keep in mind that the attentional input consisted only of two additional values (which the network will interpret as the center of attention), and that

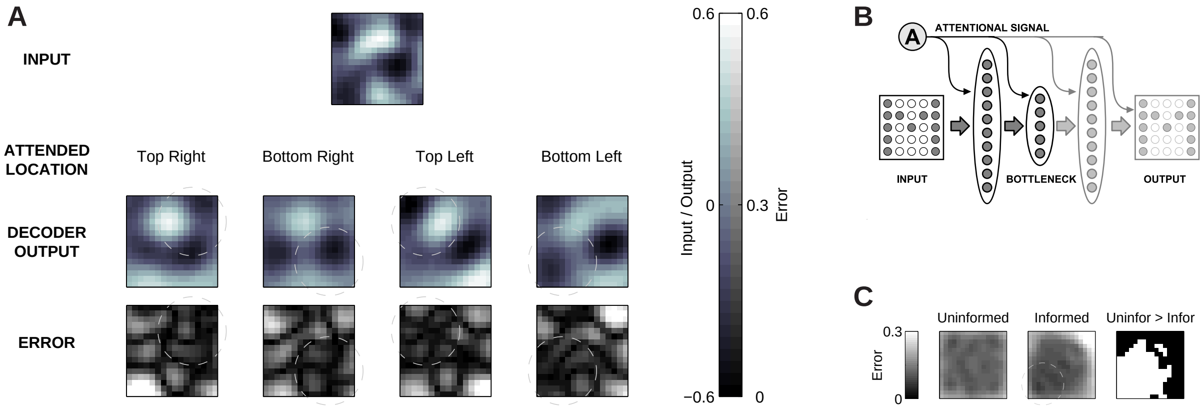

Figure 1: Reallocation of resources was observed when attentional signal changed. (**A**) Example of reconstruction of a single input pattern and four different attentional states as indicated by the dashed circles. Error is lower for attended locations. (**B**) Structure of the feedforward network used in the simulations. For each unit, attentional inputs (after training) are indistinguishable from sensory inputs. (**C**) Mean over patterns of position-specific error for one attentional state, compared against that of a system with no attentional signal. The white region in the image on the right indicates lower error when the system makes use of the attentional signal.

these signals entered each layer the same way as feedforward (sensory) inputs. The regions represented by the dashed circles were not explicitly represented by any signal in the network during this simulation (only during optimization, through the coefficients c_k in Equation (2)). As expected, lower error was achieved for those regions closer to the center of attention.

After confirming that the system's processing was dependent on attentional state, we asked how these results compared to those from a classical model in which there is no informing signal. Fig. 1C compares the mean reconstruction performance of a system which ignores the informing signal with a system attending to the bottom-left corner of the input image. The difference shows that reconstruction was better for attended regions but worse for unattended ones.

Preferred stimuli of bottleneck units was dependent on attentional state

Results presented in this section concern the units from the bottleneck layer. Activity from these units represent a lossy encoding of the input image. Fig. 2 shows the stimulus that maximally drove each of these units at different attentional states, as indicated by the dashed circles. Results are clearly dependent on attentional state, with preferred stimuli containing sharper edges in regions closer to the center of attention.

Activity of bottleneck units was modulated by non-sensory signals

How was the activity in the bottleneck neurons affected by the attentional signals? Fig. 3 shows maps of activity created by fixing the stimulus and moving the center of attention around the image. The intensity at each point in the map represents the activity of one unit in the bottleneck, normalized in the range of activities observed for that particular stimulus. The bars between the two rows of maps represent the range of activities achieved in each condition. The pattern representing activity modulation was clearly different for each unit. Furthermore, the modulation of a unit's activity was dependent on the stimulus. For example, for unit U1, higher changes in activity occurred when attention shifts from left to right or from top to bottom, depending on the stimulus. This effect occurred even when, in the case of U1,

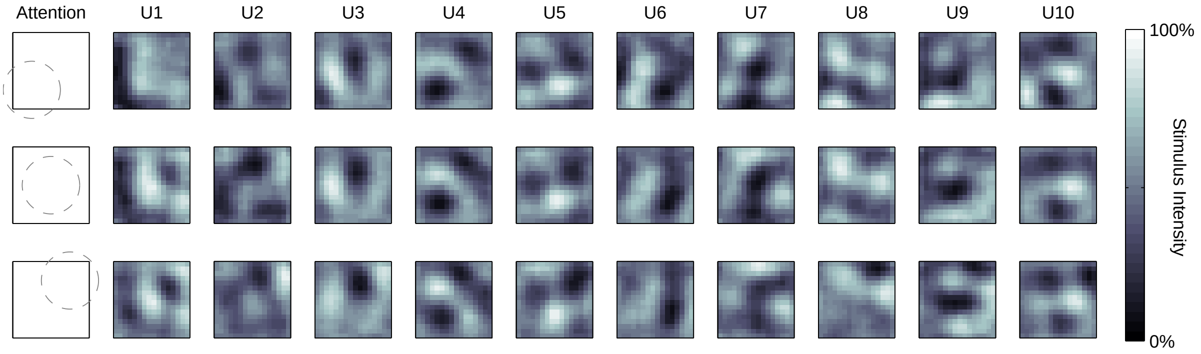

Figure 2: Preferred stimuli were dependent on attentional state. Color-coded images represent the preferred stimulus for each unit in the bottleneck layer (columns) for three attentional states (rows) as indicated by the dashed circles. Higher contrast was observed on attended regions.

the activity ranges were very similar for the two conditions.

The model reproduced measurements from visual cortex

We related these findings to experimentally observed modulation of neural activity during selective attention tasks. First, we replicated the analysis presented in Treue and Maunsell (1999), in which combinations of preferred and anti-preferred stimuli were presented inside the receptive field (RF) of a cell and activity was measured as attention changed between these two regions. Fig. 4A shows the response from a neuron in the medial superior temporal (MST) area. The preferred stimulus for this neuron was a pattern of dots moving in one direction, indicated by the upward-pointing arrow. The attended stimulus is indicated by the dashed ellipse. The histograms in gray show the firing rate of the cell as a function of time, and mean responses are indicated by solid horizontal lines for each condition. Overlaid, we show the response of one neuron from our model (dotted lines) under similar conditions. The stimuli consisted of combinations of the left and right halves from the preferred and anti-preferred stimuli, calculated with attention directed to the center of the input space. The maximal firing rate for the model neuron was set to 50 spikes/sec to obtain comparable magnitudes. The modulation of activity of the model neuron matched that of the visual neuron. In particular, when attention was shifted from preferred to non-preferred features of the same stimulus, activity decreased dramatically.

The scatter plot in Fig. 4B shows the firing rates for neurons in areas MT and MST when attention was directed toward the preferred stimulus (y axis) versus the firing rates obtained while attending the anti-preferred stimulus (x axis). Points above the diagonal indicate higher activity when attending the preferred stimulus. Values for all model units, indicated by stars, fall above the diagonal and within the range of experimentally observed values. For this plot, the range of the $s(\cdot)$ function was scaled and shifted to 0–100 spikes/sec.

Further exploration of these effects is shown in Fig. 5. Responses to the four combinations of preferred and anti-preferred half-stimuli are presented, first for a single unit, and then for all units in the bottleneck layer. Fig. 5A shows the stimuli and corresponding attention maps for unit U2. The stimuli consisted of combinations of the left and right halves from the preferred and anti-preferred stimuli. Attention maps showed a clear change in the activity of the unit as attention was shifted from right to left. The simulation also showed that changes in features far from the attended region have a smaller effect on activity than changes presented at the attended location (compare dark bars of Fig. 5B). As shown in Fig. 4, when attention

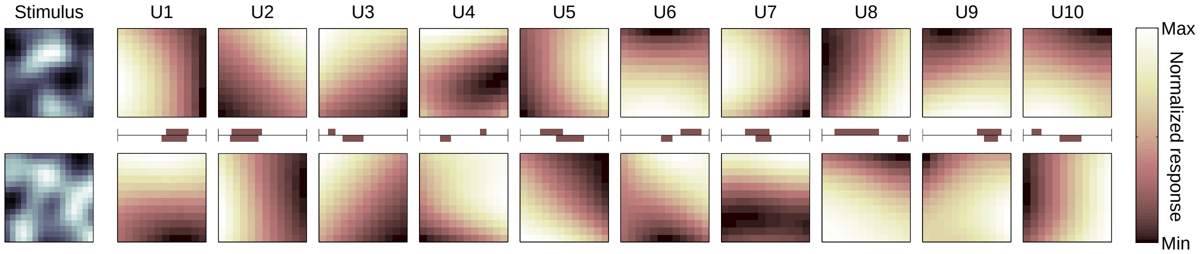

Figure 3: Activity was modulated by the attentional signal. Color-coded images represent the activity of each bottleneck unit for a fixed stimulus as the center of attention is directed to different regions of the input space. Two different stimuli (top and bottom rows) are used for comparison. Maps are scaled according to the range of activities observed for that particular stimulus. The bars between the two rows display the range of activity for each of the two conditions with respect to the absolute limits of the activity of the units.

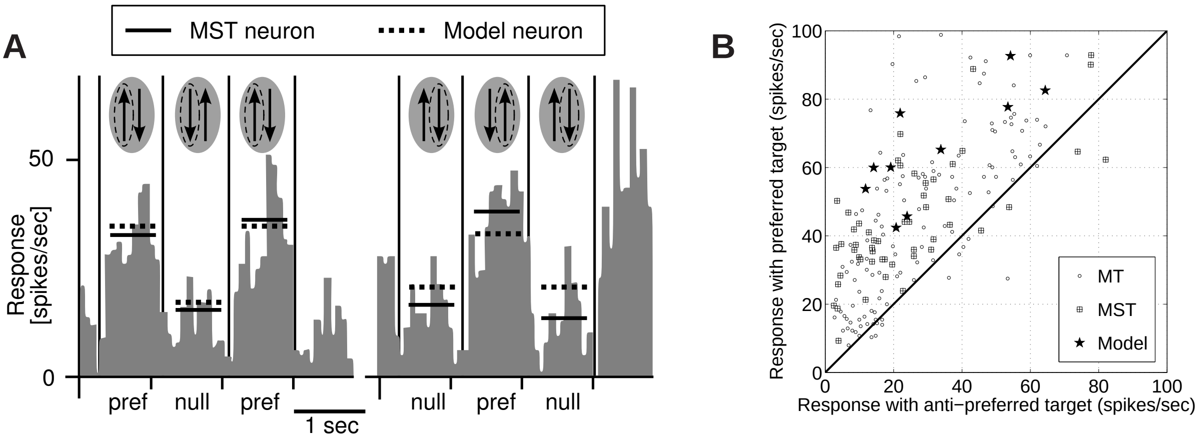

Figure 4: Activity modulation matched experimental measurements from area MST. (**A**) Neuronal response to two stimuli inside the receptive field of the cell, one preferred (arrow up) and one anti-preferred (arrow down). Attentional focus is indicated by a dashed ellipse. The histograms in gray show the firing rate of the cell as a function of time. Mean responses are indicated by solid horizontal lines (MST cell) and dotted lines (model unit) for each condition. (**B**) Scatter plot showing the response to anti-preferred stimuli *vs.* the response to preferred stimuli for each unit. Points above the diagonal indicate higher activity when attending to the preferred stimulus. (Reproduced from Treue and Maunsell, 1999)

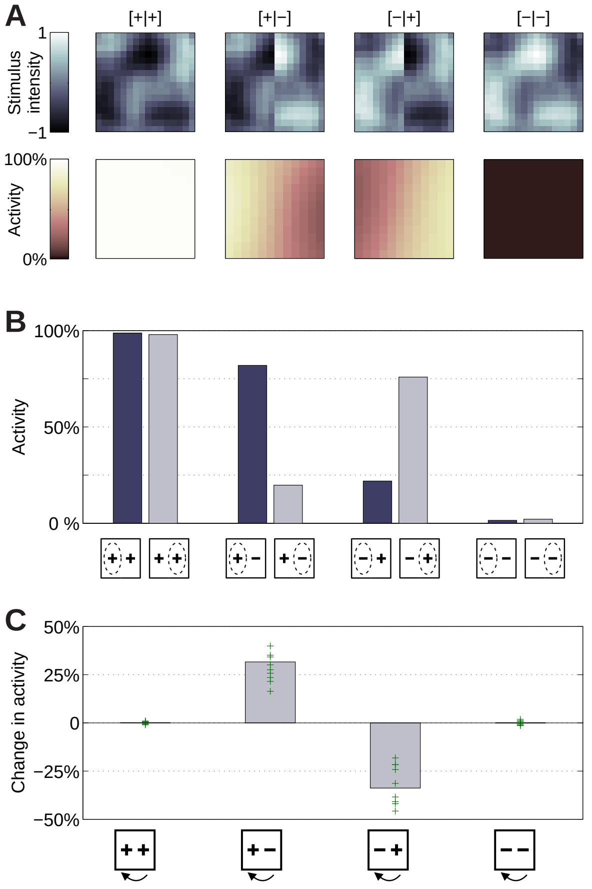

Figure 5: All model units showed similar activity modulation. (**A**) Combination of preferred (**+**) and anti-preferred (**–**) stimuli for unit U2 (top) and attentional modulation of activity in U2 for these inputs (bottom). (**B**) Comparison of activity in U2 for two attentional states as indicated by the dashed ellipses. (**C**) Changes in activity for each unit in the bottleneck as attention is shifted from right to left. Bars show means across units.

was shifted from preferred to non-preferred features of the same stimulus, activity decreased dramatically. In contrast, the effect of attention when both halves were preferred or anti-preferred was very small. These observations were consistent for all units (Fig. 5C).

The model was also compared to experiments in which the response of cells in area V4 were measured for four attentional conditions, while a bar of fixed orientation was displaced inside the receptive field of the cell (Connor et al., 1996, 1997). Fig. 6A shows the response of one V4 cell. These plots contain various features that are common in attentional modulation:

(1) The response of the cell for a fixed stimulus depends on the attended location.

(2) The stimulus that elicits the strongest response depends on attention.

(3) The cell's response when attention is fixed depends on the stimulus position.

(4) This dependence on stimulus position differs between attentional states—for instance, the left and right panels in Fig. 6A display different trends as the stimulus position is changed.

Fig. 6B shows the results from a model neuron under similar conditions. In this case, the input image is composed of a non-preferred stimulus with a small region belonging to the preferred stimulus at different positions, indicated by the numbers 1–5. Attention is directed to the borders of the image, as indicated by the circles. The model exhibited all features observed in the experimental data.

Two shift indexes were also calculated for each neuron, after Connor et al. (1997). The *fractional shift* measures the proportion of total response that shifted from one side of the RF to the other when attention is shifted. This index is bounded between -1 and 1, which a positive value indicating shifts in the direction of attention. Connor et al. (1997) reports mean values of 0.16 or 0.26 depending on whether 5 or 7 bar positions were in use. All our model neurons had a positive fractional shift, with a mean value of 0.22. The second index, the *peak shift*, measures the distance between positions generating maximum responses. Mean experimental reported values were 10% or 25% of the RF size, depending on whether 5 or 7 bar positions were in use. Our model neurons had non-negative peak shifts with a mean value of 25% of the total position variation.

Magnitude of attentional modulation depended on capacity

The mean modulation of activity of bottleneck units in the model as attention was shifted from left to right is shown in Fig. 7, with panel A showing how this varies with the number of bottleneck units, and panel C showing

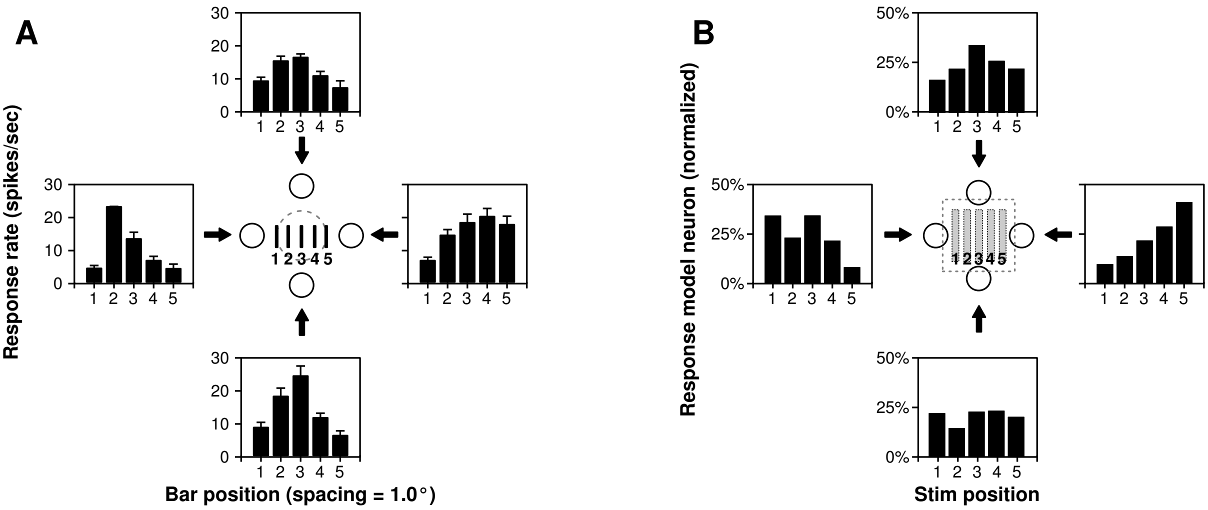

Figure 6: Activity modulation matched experimental measurements from area V4. (**A**) Response of one V4 neuron to a bar stimulus placed at five different positions inside the receptive field, as indicated by a dashed circle. Attention was directed to one of the four circles outside the receptive field. (Reproduced from Connor et al., 1996). (**B**) Response of one model neuron as a region of a non-preferred stimulus is replaced by the preferred stimulus in five different locations. Attention is directed to the border of the input space as indicated by the circles.

how this varies with amount of injected noise. Gray open circles correspond to the absolute value of the modulation averaged over 1000 random test patterns, for each unit in the bottleneck. The solid circles show the mean across units, with error bars indicating the standard error. Panels B and D show the corresponding reconstruction error for each set of parameters, plotting the mean errors for the attended versus unattended halves of the sensory input. Note that noise was added to the bottleneck units only during optimization, and for this reason changes in the modulation magnitude cannot be attributed to noise in the measurements.

The magnitude of the modulation decreased as the number of bottleneck units increased. As expected, errors in the ignored region were higher than those in the attended region, but decreased as the number of bottleneck units was increased, gradually closing the gap. As the noise level in the bottleneck units was increased, the magnitude of the modulation and of the reconstruction error both increased. The error for the attended region increased faster than for the unattended region. These two plots display an abrupt transition as the noise level becomes very high: the trend of the activity modulation, as well as the error values, changes, with the error values for the attended and unattended regions becoming equal.

4 Discussion

The above results show that, when a neural system is optimized to encode its inputs with changing fidelity requirements, the code developed exhibits modulatory phenomena of the sort that has been observed in the visual systems of animals engaged in selective attention tasks. These modulatory phenomena, which constitute a reallocation of resources, emerge even when the encoder is a very simple homogeneous feedforward neural model.

Limitations of the Simulation

The simulations incorporated a number of simplifications, most of which were made for ease of exposition or computational efficiency.

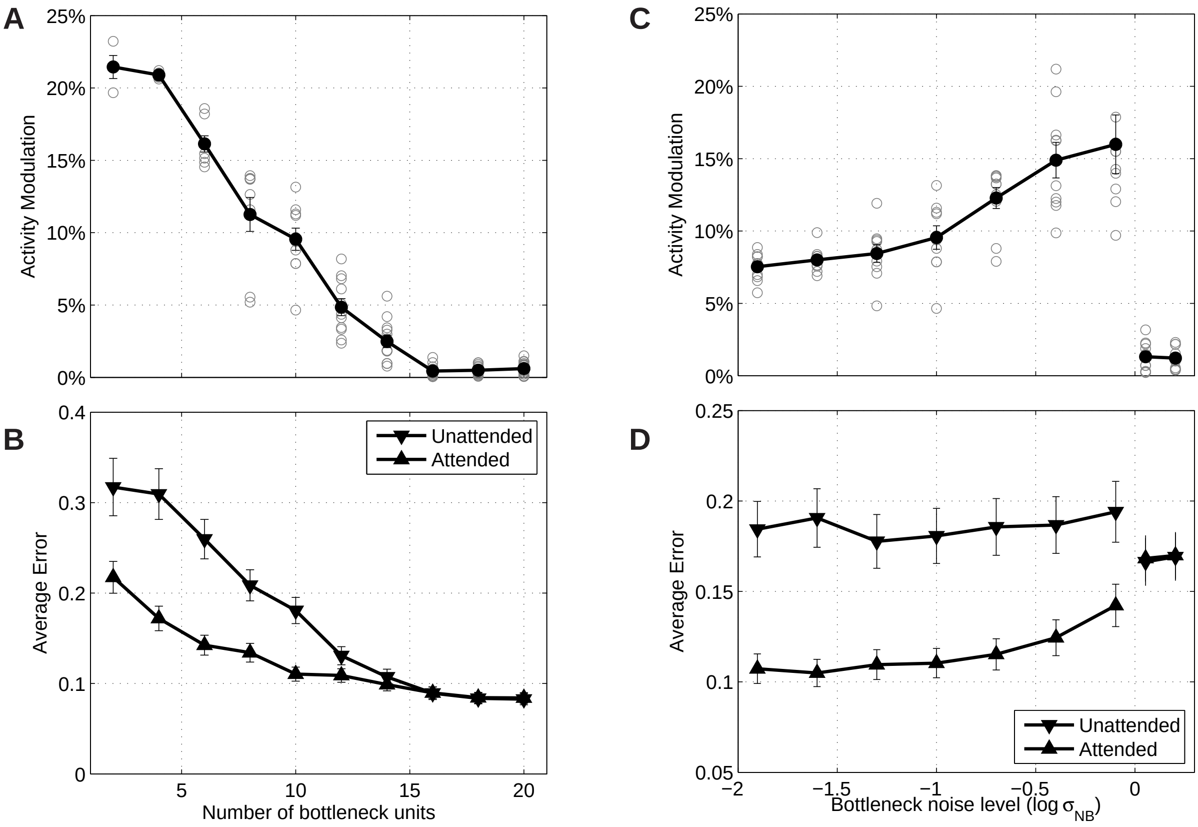

Figure 7: Magnitude of attentional modulation depends on system capacity. (**A**) Modulation of activity in response to random stimuli as attention was shifted from left to right, for networks with different numbers of units in the bottleneck. Open gray circles represent the mean modulation for each unit, while solid circles show the mean over all units with bars indicating the standard error. (**B**) Mean reconstruction error for the attended and unattended sides. (**C**,**D**) Same as (A,B) but using 10 bottleneck units and instead varying the amount of noise injected into the bottleneck units during training.

(i) During optimization, the penalty function was set to a single attentional spotlight. This is not a requirement of the general model, and other functions could be used to define the performance demands. For example, non-spatial goals could be incorporated by requiring higher fidelity reconstruction of some particular feature regardless of its spatial location.

(ii) To allow convenient visual display, simulated stimuli were visual patterns. Efficient representation of stimuli and attentional modulation phenomena are present in many (if not all) modalities, and the model explored here could be applied equally well to non-visual and non-spatial modalities.

(iii) The simulation allowed synapses from a single neuron to be both excitatory and inhibitory. This is not a common feature in biological neural systems, where a combination of excitatory and inhibitory neurons could achieve the same functionality.

(iv) The simulation is limited to explorations inside the receptive field of a cell. Furthermore, as a consequence of the simulation's simplicity, it cannot make detailed predictions about the form of the RFs. Extended and more complicated models would be necessary to predict attentional modulation of realistic RFs. For instance, incorporating sparsity constraints into the optimization procedure should give rise to more realistic localized RFs (Olshausen and Field, 1996), which would allow the prediction of attentional modulation of RF shape and location.

(v) The capacity in the bottleneck was restricted by limiting the number of neurons, with each neuron's capacity limited by an intrinsic noise term. Given that primary sensory cortex typically has greater than $10^3\times$ more neurons than the sensory sheet it represents, e.g., in humans, 10^6 retinal ganglion cells versus over 10^9 V1 pyramidal neurons, one is led to question the biological relevance of this capacity restriction. It is important to note that this is not an issue with the model of attention per se, but rather with the optimal coding framework it builds upon. One potential resolution of this issue is that energetic constraints may drive the nervous system to use efficient codes even in areas where there is an abundance of neurons, and even where their principal role is computation rather than coding.

(vi) The data displayed was collected after the optimization procedure had been run and the connection strengths fixed at their optimal values. In nature, we might expect this type of adaptation to be continuous, rather than being confined to a period of training.

(vii) To allow standard effective learning algorithms to be applied, the simulation was at the level of firing rates rather than spikes, and short-term plasticity was not included.

(viii) The goal of the network in the simulations was to find an efficient representation for the stimulus. Biological networks likely transform stimuli not to merely compressed representations, but to representations that serve to inform future actions.

The limitations presented above are mostly specific to the current simulation rather than to the general model proposed in this paper, or to the predictions it makes. Stimulus-driven attentional phenomena, and the undefined origin of the attentional signal, are some of the limitations that elaborations of the model might address. Despite these simplifications, the simulation exhibited many phenomena observed experimentally, and makes novel concrete testable predictions about the modulation of neural activity by attentional processes. The fact that the model, even with these simplifications, accounts for a broad range of previous measurements and makes novel robust predictions is, to our mind, a strength rather than a weakness.

Consistency with electrophysiological data

The predictions of this model are consistent with observations from electrophysiological recordings in which the output of neurons in the visual system are modulated by the attentional state of the subject. This modulation suggests that the characterization of neural response should go beyond the traditional stimulus-response dependency, usually represented as a receptive field or tuning curve.

Treue and Maunsell (1999) showed that when combining preferred and non-preferred moving stimuli inside the receptive field of a cell in monkey area MST, the activity of the cell was higher when attention was directed to the region containing the preferred direction of motion. These results are exhibited by the model (Fig. 4 and 5).

Results from the model are also consistent with those of Connor et al. (1997). Experimental measurements of the modulation of neuronal response for different sets of stimuli and attentional conditions qualitatively matched those from model neurons. Furthermore, when setting the maximum firing rate parameter to biologically plausible values, the model produced modulations that lie in the ranges observed experimentally. We would make one observation regarding the shift indices of Connor et al. (1997): it may be misleading to interpret the partial shift values merely as a tendency to have increased activity when a bar is presented at the attended edge of the RF, as compared to the activity when a bar is presented at the opposite edge of the RF—a property exhibited by the model of Rao (2005). For instance, even when all units show positive indices, some may show higher activity for a bar in position 1 compared to position 5 in both (left and right) attentional conditions. What the index measures is the *proportion* of activity that shifts when attention changes.

Implicit in the measurements of Treue and Maunsell (1999) and Connor et al. (1997) is the fact that the attentional modulation is dependent on the stimulus, *e.g.*, attention moving from left to right inside the RF of a cell will have an increasing or decreasing effect on the neuron's firing rate depending on the relative location of preferred and non-preferred features. Examples of this dependency in our simulation are shown in Fig. 3. This phenomenon implies that attentional modulation cannot be fully characterized using a single stimulus.

Treue and Maunsell (1999) also observed that directing attention towards the receptive field of a cell can both enhance and reduce responses, *e.g.*, when the animal was attending to the anti-preferred of two directions in the receptive field, the response was below the one evoked by the same stimulation when attention was directed outside the receptive field. While this phenomenon could not be tested directly using our simulation, our results suggest a clear reduction in activity when attention is directed to the anti-preferred region compared to other regions (Fig. 5). This is not obvious from models in which attention simply amplifies activity for attended locations. Furthermore, the model predicts that stimulus changes in the attended location generally produce a higher change in activity than stimulus changes in unattended locations (Fig. 5B).

Results from our model are consistent with other experimental results not analyzed in detail in this paper. For instance, the response of cells in areas V2 and V4 is attenuated when placing a non-preferred feature inside the receptive field when compared to a preferred feature alone. If attention is directed to the preferred feature, activity is restored (Reynolds et al., 1999). Our model exhibits the same phenomenon, as suggested by Fig. 5. Another example relates to the type of modulation attentional signals produce on neural activity. Preliminary results suggest that this model can display apparent multiplicative effects without explicit multiplicative interactions. A more complex model that exhibits localized RFs, as discussed above, would be necessary to obtain conditions comparable to those from the electrophysiological recordings discussed in (McAdams and Maunsell, 1999).

Comparison to other modelling approaches

Selective attention researchers have suggested a wide variety of models with different predictive power, most of them presenting accounts of behavioral phenomena rather than explicit predictions on the modulation of neural activity (Mozer and Sitton, 1998; Deco and Zihl, 2001; Heinke and Humphreys, 2003).

An influential early modelling study that made concrete predictions regarding attentional changes of neural activity was developed in Olshausen et al. (1993). In that study, control neurons dynamically modified the synaptic strengths of the connections in a model network of the ventral visual pathway. The network selectively routed information into higher cortical areas producing invariant representation of visual objects. This model predicted changes in position and size of receptive fields as attention was shifted or rescaled. These phenomena are partially supported by results from Connor et al. (1997). Their model also qualitatively matches modulation effects observed by Moran and Desimone (1985) with stimuli inside and outside the classical receptive field of V4 neurons. In comparison to our model, in which attentional modulation emerges from the nonlinearity of the units and general objective of the network, their model obtained modulatory effects by explicitly modulating the synaptic strengths of the connections. Their model also used a spatially localized connectivity pattern which gave rise to localized RFs, thus allowing for comparison of attention directed inside versus outside the RF.

More recent studies incorporate principles of statistical inference into models of attention. For instance, Yu and Dayan (2005) and Rao (2005) present networks that implement Bayesian integration of sensory inputs and priors, and which replicate behavioral as well as electrophysiological measurements. In these studies, spatial attention is equated to prior information on the location of the features of interest. The Bayesian inference approach to modeling attention should be regarded as complementary to that taken here. The transformations performed by the model units in the present work are defined by the solution to an optimal coding problem; and under certain conditions, these computations would be equivalent to those in inference-based networks. In fact coding, statistical modeling of distributions, and inference from partial data are, mathematically speaking, very closely related.

Mechanisms that subserve attentional modulation

Local gain modulation, a common tool in mechanistic models of attentional modulation, is not necessarily in opposition with our approach. What we suggest is that events traditionally described as local gain modulation subserve a global reallocation of resources, which is the strategy the nervous system has evolved for approaching optimal performance given its constraints.

The mechanisms of attentional modulation have been traditionally posited to be changes in synaptic efficacy or modulation of presynaptic terminals (Olshausen et al., 1993). Here we show that the overall effects of this modulation can appear in a network of saturating units without any changes in synaptic strength, being controlled only by the activity itself. This is consistent with the notion that a network optimized for efficient coding under shifting fidelity requirements will integrate whatever information is available about the current requirements by pressing into service any available mechanism.

In other words, it is possible that there is no way to anatomically distinguish between neuronal mechanisms that support coding per-se from mechanisms that support shifts in coding. While it could be argued that evidence of multiple sites of integration in cortical neurons (Larkum et al., 1999) is inconsistent with this idea, the results above show that attentional modulation of the neural code is not sufficient to explain the functional role of inputs at different cortical layers, instead suggesting that these mechanisms may be more important for learning or other processes.

Main Predictions

The fact that our simulation shows modulation effects consistent with physiological recordings suggests that we should not necessarily expect explicit gating circuitry in neural systems responsible for attentional phenomena. Furthermore, the informing signals do not have to explicitly represent the "attentional space", *i.e.*, a spatial attention effect is not necessarily mediated by a topographic input.

Our model strongly predicts that a neuron's "preferred stimulus" will depend on attentional state. Moreover, the behavior of a single neuron in this model cannot be well characterized by measurements of attentional modulation of only a single sensory stimulus. This prediction is consistent with experimental results discussed above, but it should be possible to test it more explicitly using currently available experimental techniques.

The model also suggests that stronger modulations are expected when the complexity of the input grows, relative to the capacity of the system.

5 Conclusion

The model presented here accounts for attentional modulation of neural response in a framework that includes both attention and receptive field formation, and as a consequence of an underlying normative principle (optimal coding) rather than by tuning a complex special-purpose architecture. The model shows that reallocation of resources can emerge even in a simple feedforward network, and challenges the traditional characterization of neural activity. These results are consistent with the notion that attentional modulation is not, at its root, due to specific local architectural features, but is rather a ubiquitous phenomenon to be expected in any system with shifting fidelity requirements.

Acknowledgments

Supported by US NSF PCASE 97-02-311, the MIND Institute, an equipment grant from Intel, a gift from the NEC Research Institute, Science Foundation Ireland grant 00/PI.1/C067, and Higher Education Authority of Ireland (An tÚdarás Um Ard-Oideachas.)

References

J. J. Atick. Could information theory provide an ecological theory of sensory processing? *Network: Comput. in Neural Sys.*, 3(2):213–51, May 1992.

C. E. Connor, J. L. Gallant, D. C. Preddie, and D. C. Van Essen. Responses in area V4 depend on the spatial relationship between stimulus and attention. *J Neurophysiol*, 75 (3):1306–8, 1996.

C. E. Connor, D. C. Preddie, J. L. Gallant, and D. C. Van Essen. Spatial attention effects in macaque area V4. *J Neurosci*, 17(9):3201–3214, 1997.

M. Corbetta and G. L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci*, 3(3):201–215, 2002.

P. Dayan and R. Zemel. Statistical models and sensory attention. In *Proceedings of the Ninth International Conference on Artificial Neural Networks*, 1999.

P. Dayan, S. Kakade, and P. R. Montague. Learning and selective attention. *Nature Neuroscience*, 3:1218–1223, 2000.

G. Deco and J. Zihl. A neurodynamical model of visual attention: feedback enhancement of spatial resolution in a hierarchical system. *Computational Neuroscience*, 10(3): 231–253, 2001.

C. Downing. Expectancy and visual-spatial attention: effects on perceptual quality. *J Exp Psychol Hum Percept Perform*, 14(2): 188–202, 1988.

D. Heinke and G. W. Humphreys. Attention, spatial representation, and visual neglect: simulating emergent attention and spatial memory in the selective attention for identification model (SAIM). *Psychol Rev*, 110 (1):29–87, 2003.

H. J. Heinze, S. J. Luck, T. F. Münte, A. Gös, G. R. Mangun, and S. A. Hillyard. Attention to adjacent and separate positions in space: an electrophysiological analysis. *Percept Psychophys*, 56(1):42–52, 1994.

J. B. Hopfinger, M. H. Buonocore, and G. R. Mangun. The neural mechanisms of top-down attentional control. *Nature Neuroscience*, 3(3):284–291, 2000.

S. Jaramillo and B. A. Pearlmutter. A normative model of attention: Receptive field modulation. *Neurocomputing*, 58-60:613–8, 2004.

M. E. Larkum, J. J. Zhu, and B. Sakmann. A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature*, 398(6725):338–41, 1999.

S. J. Luck, L. Chelazzi, S. A. Hillyard, and R. Desimone. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J Neurophysiol*, 77(1):24–42, 1997.

C. J. McAdams and J. H. R. Maunsell. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J Neurosci*, 19(1):431–441, 1999.

J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229(4715):782–784, 1985.

M. C. Mozer and M. Sitton. Computational modeling of spatial attention. In H. Pashler, editor, *Attention*, pages 341–393. Psychology Press, 1998.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

B. A. Olshausen, C. H. Anderson, and D. C. Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J Neurosci*, 13(11):4700–4719, 1993.

R. Rao. Bayesian inference and attentional modulation in the visual cortex. *Neuroreport*, 16(16):1843–8, 2005.

J. H. Reynolds, L. Chelazzi, and R. Desimone. Competitive mechanisms subserve attention in macaque areas V2 and V4. *J Neurosci*, 19(5):1736–1753, 1999.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back–propagating errors. *Nature*, 323:533–6, 1986.

S. Treue and J. H. R. Maunsell. Effects of attention on the processing of motion in macaque middle temporal and medial superior temporal visual cortical areas. *J Neurosci*, 19(17):7591–7602, 1999.

A. Yu and P. Dayan. Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–92, 2005.