A Normative Model of Attention: Receptive Field Modulation

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

Abstract

Normative models, which explain structure as being optimized to perform some function well, dominate theoretical biology. Such models inform our understanding of everything from the size of parts of the circulatory system to the thickness of a whale's blubber to receptor density on an *E. coli*. Such models have recently entered neuroscience, finding particularly fruitful application in the study of low-level sensory processing. Here we introduce a normative model of top-down attentional modulation in which a top-down attentional signal (whose origin is outside the scope of the model) breaks a symmetry implicit in the standard information-theoretic account of low-level sensory receptive field formation by modulating the tradeoffs in transmission fidelity of various characteristics of the pattern being encoded. Under suitable conditions this causes the internal representation assigned to an input pattern to shift as the top-down attentional signal changes. We instantiate this general model in a simple network that *learns* a covert attentional response modulated by a top-down attentional signal. The network is quite general, consisting of an autoencoder with a bottleneck where the signals through the bottleneck constitute the compressed encoding of the sensory input. The encoding (compression) and decoding (pullout) processes are both informed by the top-down attentional signal, allowing the compressed representation to be modulated when such modulation is advantageous. Both the encoding and decoding networks are generic function approximators (MLPs) which learn their mappings without any domain-related bias or structure. The only information they receive concerning the semantics of the top-down attentional signal is from the optimization criterion, which for the specific visual phenomena being modeled here penalizes the system more heavily for errors made near an attentional spotlight. After detailing the architecture, we explore similarities between the modulation of neural activity in our network and that in the visual systems of animals measured during covert visual attention tasks. This model provides a simple, general, and predictive account of top-down modulation of receptive fields.

1 Introduction

The limited capacity for processing information and the ability to filter out unwanted information are the basic phenomena that define attention (Desimone and Duncan, 1995). Attention has been often seen as an information-processing bottleneck that allows only part of the incoming sensory information to reach short-term memory and awareness. Attention is usually described as being composed of two processes. The first, charged with recognizing salient patterns, is primarily bottom-up (Itti and Koch, 2001). The second is seen as task-dependent, and is most probably controlled from higher areas which connect back into early sensory areas (top-down). In this paper we focus on the latter.

Selective attention can occur covertly, *i.e.* without overt movements of the head or sensory organs. Covert attention has been shown to improve discriminability and to accelerate the rate of information processing (Carrasco and McElree, 2001).

Attentional mechanisms are present in all modalities, and are not always spatial but can be feature- or object-based (Motter, 1994; McAdams and Maunsell, 2000; Treue and Martínez-Trujillo, 1999; O'Craven et al., 1999). Desimone and Duncan (1995) suggest that attentional mechanisms for object and spatial selection may work in a similar fashion, differing only in the source and nature of the selection template

Different brain areas have been shown to be involved in preparing and applying goal-directed (top-down) selection for stimuli and responses. Barceló et al. (2000) present evidence on prefrontal cortex regulation of neuronal activity in extrastriate cortex. Other imaging studies (Corbetta and Shulman, 2002; Hopfinger et al., 2000) suggest that parietal and temporal areas are also involved in

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

attentional mechanisms. Measurements from visual cortex suggest a retinotopic mapping of attention-related activation (Brefczynski and DeYoe, 1999).

Physiological studies have shown that when two stimuli are presented simultaneously inside a cell's receptive field, the cell's response is strongly influenced by which of the two stimuli was attended (Moran and Desimone, 1985; Luck et al., 1997; Treue and Maunsell, 1999; Reynolds et al., 1999; Posner and Gilbert, 1999). Other studies have focused on the modulation of the receptive fields in the visual cortex (Connor et al., 1997; Ben Hamed et al., 2002). But the fashion in which this modulation occurs remains a subject of debate (Treue and Maunsell, 1999; McAdams and Maunsell, 1999; Reynolds et al., 2000; Chawla et al., 1999; Fries et al., 2001; Steinmetz et al., 2000).

Previous computational models of top-down attention use gating mechanisms or connection weight modulation to implement selective attention (Hinton and Lang, 1985; Olshausen et al., 1993; van de Laar et al., 1997; Deco and Zihl, 2001; Heinke and Humphreys, 2003), and the particulars of the modulation is thus built into those models. This paper introduces a new class of model, in which the attentional signal is presented to the processing layers in the same fashion as is the sensory input, and the system learns to assign resources to different parts of the stimulus, and to modulate this assignment according to the attentional signal, without any special structure or architectural bias. The only information concerning the semantics of the attentional signal comes from the error measure, which for the specific visual phenomena being modeled here penalizes the system more heavily for errors made near the center of an attentional spotlight.[1]

We start by describing the architecture of our model, followed by a demonstration of the learned modulation of unit activation. This modulation qualitatively matches the results found in animals during covert visual attention tasks (Moran and Desimone, 1985; Treue and Maunsell, 1999).

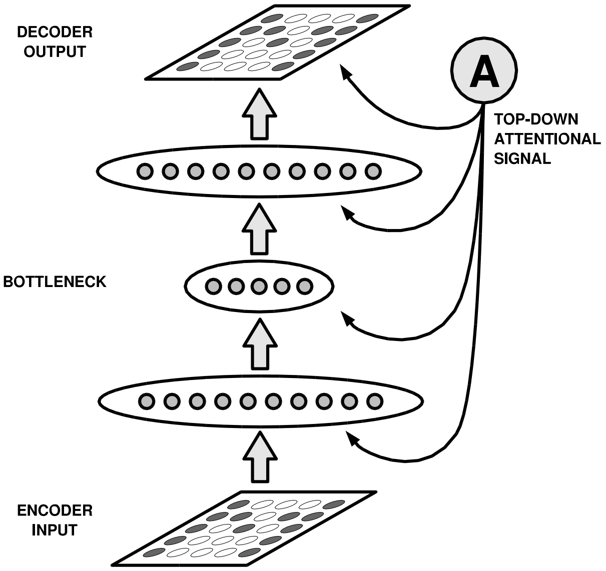

Figure 1: Network architecture. The network contains five layers forming an encoder/decoder system with a bottleneck. Each layer has an additional attentional signal.

2 Methods

2.1 Network architecture

The implemented system is an auto-associative network composed of five layers, of which the third is a bottleneck, as shown in Figure 1. The bottom layers can be seen as encoding the input signal, and the top layers as decoding it. The middle layer or bottleneck will represent the input pattern using fewer units[2] than its original representation in the input layer.

The units between one layer and the next are fully connected and they receive an additional input denoted here as *top-down attentional signal* (see Figure 1). The number of units per layer is 256–20–10–20–256, where the input and output layers are treated as 16×16 grids for display purposes. The attentional signal consists of a two-element vector representing the center of the attention mask. The units use hyperbolic tangent activation functions which saturate at ±1.71, and the attentional signals representing the (x,y) coordinates of the center of the attentional spotlight were scaled and shifted into the range ±1.

[1] The framework admits more abstract attentional goals, such as low reconstruction error for faces in an image, or retention of features to distinguish alphabetical characters, or retention of acoustic features of one source with no penalty for discarding those of an interfering acoustic source. Using such more sophisticated goals may allow the model to account for phenomena such as popout, to predict new sorts of attentional modulation in response to changes in higher-level goals, and to be applied in non-visual modalities. But for the sake of simplicity, and in order to compare the model to measured data, this paper confines itself to a simple "spotlight" attentional goal.

[2] Fewer units is not technically necessary, as other means (such as injected noise, an activation penalty in the optimization criterion, or even the information bottleneck (Bialek et al., 2001)) can also serve to limit the capacity of the bottleneck.

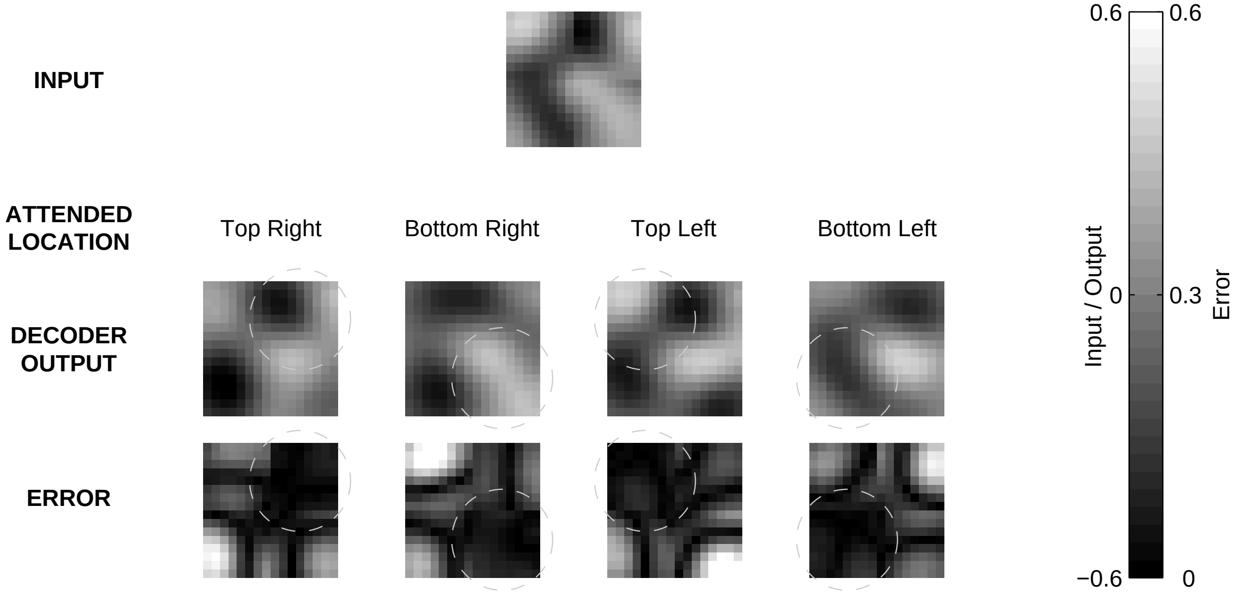

Figure 2: Reconstruction example. The top image corresponds to the input stimulus. The following rows of images show the output and decoding error respectively. Each column corresponds to a different value of the attentional signal, *i.e.* a different location of the attentional spotlight. The dashed circles indicate where attention was directed in each condition. The error is calculated as the absolute pixel intensity difference between the input and the output patterns. Darker tones correspond to lower error.

2.2 Training

The network was regarded as a single system, with the encoder and decoder portions jointly optimized to minimize

$$E(\mathbf{p}) = \sum_i c_i(\mathbf{p})\,(y_i(\mathbf{p}) - d_i(\mathbf{p}))^2 \qquad (1)$$

where $c_i(\mathbf{p})$ is the intensity of the attentional spotlight at location i, $y_i(\mathbf{p})$ is the output of the network at location i for input $\mathbf{p}$, $d_i(\mathbf{p})$ is the desired output at location i which is of course the same as the input at location i, and $\mathbf{p}$ represents a tuple holding the complete input to the system, *i.e.* the input pattern as well as the top-down attentional signal.

The gradient was calculated using backpropagation (Rumelhart et al., 1986) and optimization used simple 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

$$c_i(\mathbf{p}) = \frac{1}{1 + k^2 ||i - a(\mathbf{p})||^2} \qquad (2)$$

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

2.3 Training Set

The 2000-element training set consisted of 16×16 pixel images, with the pixels being zero mean and having stdev $\sigma = 1/3$. The images were created by convolving (filtering) white Gaussian noise images with a rotationally symmetric 2D Gaussian with $\sigma_{filter} = 2$. Edge effects were avoided by extracting only the 16×16 center of the resulting image. These images were later scaled to have the desired stdev. The center of the attentional mask was drawn independently of the input image, and uniformly distributed within the input image.

2.4 Controls

It is important to note that with real-valued units in the bottleneck the capacity of the system would be theoretically infinite, assuming sufficiently sophisticated encoding and decoding processes. To explicitly control capacity, zero-mean Gaussian noise with standard deviation 0.1 was added to each bottleneck unit's total input during training.

Moreover, we would like to test that the system is not just degrading its encoding/decoding performance in those locations where attention is low, but is devoting the resources made available from degraded accuracy outside the attentional spotlight to increasing the accuracy within

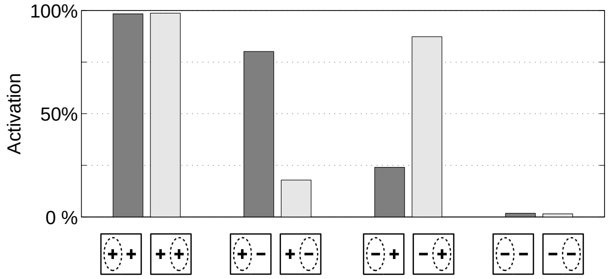

Figure 3: Activation of one bottleneck unit when attention is directed to left or right. The square under each bar indicates the stimulus. Stimuli were created by combining the left and right halves of effective/non-effective stimuli, as indicated by +/–. The dashed ellipse indicates the spotlight of attention.

the attentional spotlight. By comparing the performance of the system with a flat attentional mask to that exhibited with the peaked mask described above, we confirmed that the system is in fact reassigning resources appropriately.

3 Results

3.1 Encoding/Decoding

An example of the encoding/decoding results for a testing pattern (*i.e.* a pattern not included in the training set) is shown in Figure 2. This figure presents the output of the system when the center of the spotlight of attention is located in different corners of the image while the image itself is held constant. The dashed circles indicate the location of the attentional spotlight, but should not be interpreted as hard-edged masks. The error is calculated as the absolute pixel intensity difference between the input to the encoder network and its reconstruction by the decoder network.

3.2 Modulation of unit activation

The linear component of the encoding function for each neuron in the bottleneck was calculated using reverse correlation, *i.e.* finding the average input weighted by the activation of the unit for white-noise input (de Boer and Kuyper, 1968; Rieke et al., 1996). These values were used to define "effective" and "non-effective" stimuli for each bottleneck unit, with respect to its activation.

Images containing combinations of effective and non-effective regions were created and presented to the network. Figure 3 shows the activation of one bottleneck unit for two attentional states (right or left, as indicated by the dashed circle) and four different input images. The + and – symbols indicate which part of the image contains an effective or non-effective input, respectively.

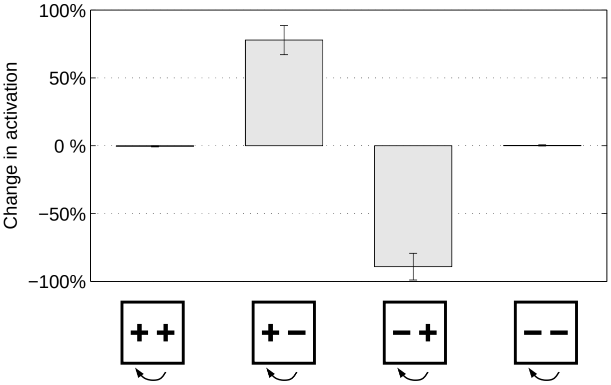

Figure 4: Modulation of unit activation by attention. The bars indicate the average change in activation over the units in the bottleneck when attention is shifted from right to left. The squares under the bars represent the stimuli, as described in Figure 3. The standard error is shown for each bar.

The difference of activation of the bottleneck units as attention is shifted from right to left is presented in Figure 4. The height of each bar represents the average over all bottleneck units (ten in our case). Standard errors are also shown.

4 Discussion

The results from the reconstruction simulation for different attentional states (Figure 2) are consistent with the hypothesis that attention assigns more resources to attended locations, thus giving better reconstruction of some features of the input stimulus. Note that in Figure 2 the error is lower inside the dashed circles, and the output pattern outside this region is different from the input stimulus.

The exploration of unit activations relates our results to physiological experiments, in which it has been shown that neural response can be modulated by an animal's attentional state, with the input remaining constant. The results from Figure 3 show a clear modulation of the activation of a unit: when the same stimulus is presented (one side effective, and the other non-effective) the activation changes dramatically depending on the top-down attentional signal. This result is common for all units in the bottleneck, as indicated by the averages and small standard errors in Figure 4.

Figure 3 also shows that changes in the stimulus in the unattended location produce smaller changes in unit activation than changes in the attended location. This can be seen by comparing the height of the second and third dark bars (attention to the left, with one half effective) with the first dark bar (attention to the left with both halves effective).

These results qualitatively match neuronal response changes found in animals under attentional modulation (Moran and Desimone, 1985; Treue and Maunsell, 1999).

McAdams and Maunsell (1999) speculates that the phenomenological similarity between the effects of attention and the effects of stimulus manipulations raises the possibility that attention involves neural mechanisms that are similar to those used in processing ascending signals from the retinas, and that cortical neurons treat retinal and attentional inputs equivalently. Our model is consistent with this notion. The relation between our normative model and other theory-driven computational models of attention, such as those suggested by Nakahara et al. (2001) or Dayan et al. (2000), remains to be explored.

One general prediction of this class of model is that an encoding system with narrow bottleneck (or rich input) will have stronger attentional modulation than a system with sufficient capacity to represent its input with greater fidelity. This might be tested by raising animals is visually rich *vs* impoverished environments and measuring differences in the magnitude of attentional modulation of receptive fields.

5 Conclusion

We presented a model that learns a covert top-down attentional mechanism, without any special structure. In this model attentional signals innervate the entire network, and each unit treats them no differently than bottom-up sensory signals. The only information received by the network about the semantics of the attentional input comes from the objective function optimized during learning. The model accounts for attentional modulation of neural response in a unified framework which accounts for both attention and receptive field formation, and as a consequence of an underlying normative principle, rather than by tuning a complex special-purpose architecture to match specific data.

The model reproduces neuronal modulation observed in physiological experiments, and can be naturally applied across tasks and across sensory modalities. The model has the potential of being extended to attentional goals where modulation would be less intuitive, such as acoustic source segregation, or feature-driven attentional goals such as priming.

References

Barceló, F., Suwazono, S., and Knight, R. T. (2000). Prefrontal modulation of visual processing in humans. *Nature Neuroscience*, 3(4):399–403.

Ben Hamed, S., Duhamel, J. R., Bremmer, F., and Graf, W. (2002). Visual receptive field modulation in the lateral intraparietal area during attentive fixation and free gaze. *Cereb Cortex*, 12(3):234–245.

Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity, and learning. *Neural Computation*, 13(11).

Brefczynski, J. A. and DeYoe, E. A. (1999). A physiological correlate of the 'spotlight' of visual attention. *Nat Neurosci*, 2(4):370–374.

Carrasco, M. and McElree, B. (2001). Covert attention accelerates the rate of visual information processing. *Proceedings of the National Academy of Sciences USA*, 98(9):5363–5367.

Chawla, D., Rees, G., and Friston, K. J. (1999). The physiological basis of attentional modulation in extrastriate visual areas. *Nature Neuroscience*, 2(7):671–676.

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

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

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

de Boer, R. and Kuyper, P. (1968). Triggered correlation. *IEEE Trans Biomed Eng*, 15(3):169–179.

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

Desimone, R. and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–222.

Fries, P., Reynolds, J. H., Rorie, A. E., and Desimone, R. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science*, 291(5508):1560–1563.

Heinke, D. and Humphreys, G. W. (2003). 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.

Hinton, G. E. and Lang, K. J. (1985). Shape recognition and illusory conjunctions. In *the Ninth International Joint Conference on Artificial Intelligence*, volume 1, pages 252–259, Los Angeles. Morgan Kaufmann.

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

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203.

Luck, S. J., Chelazzi, L., Hillyard, S. A., and Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas v1, v2, and v4 of macaque visual cortex. *J Neurophysiol*, 77(1):24–42.

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

McAdams, C. J. and Maunsell, J. H. (2000). Attention to both space and feature modulates neuronal responses in macaque area V4. *J Neurophysiology*, 83(3):1751–1755.

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

Motter, B. C. (1994). Neural correlates of attentive selection for color or luminance in extrastriate area v4. *J Neurosci*, 14(4):2178–2189.

Nakahara, H., Wu, S., and Amari, S. (2001). Attention modulation of neural tuning through peak and base rate. *Neural Computation*, 13(9):2031–2047.

O'Craven, K. M., Downing, P. E., and Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature*, 401(6753):584–587.

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

Posner, M. I. and Gilbert, C. D. (1999). Attention and primary visual cortex. *Proc Natl Acad Sci U S A*, 96(6):2585–2587.

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

Reynolds, J. H., Pasternak, T., and Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, 26(3):703–714.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., and Bialek, W. (1996). *Spikes: Exploring the Neural Code*. MIT Press. A Bradford Book.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back–propagating errors. *Nature*, 323:533–536.

Steinmetz, P. N., Roy, A., Fitzgerald, P. J., Hsiao, S. S., Johnson, K. O., and Niebur, E. (2000). Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature*, 404(6774):187–190.

Treue, S. and Martínez-Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575–579.

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

van de Laar, P., Heskes, T., and Gielen, S. (1997). Task-dependent learning of attention. *Neural Networks*, 10(6):981–992.