Awakening a Sleeping Cat: A Review of "Information Theory and the Brain" edited by R. Baddeley, P. Hancock, and P. Földiák

(to appear in *Neural Networks*, 2002)

Rajesh P. N. Rao
Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195-2350, USA
E-mail: rao@cs.washington.edu

In a seminal article published in 1954 [1], Fred Attneave provided several illustrative examples demonstrating how visual information that reaches the retina is highly redundant. He argued that an important goal of perception is to reduce this redundancy and "encode incoming information in a form more economical than that in which it impinges on the receptors" [1]. One of his most famous examples is the drawing of a sleeping cat, obtained by identifying 38 points of maximum curvature in an actual image and connecting these points appropriately with a line. The feline subject of the drawing is immediately recognizable, supporting Attneave's claim that perception involves extracting economical descriptions of objects by reducing redundancy in their raw images.

Appearing almost five decades after Attneave's inspiring article, a new book entitled *Information Theory and the Brain* reviews the progress that has been made since Attneave's cat made its somnolent appearance. The book provides an up-to-date overview of the results achieved in recent years from applying information theory to neuroscience and artificial neural networks. The book, edited by Baddeley, Hancock, and Földiák, consists of 17 chapters authored by various researchers in the field. It is derived from a 1996 conference with same name held in Newquay, England.

The scope of *Information Theory and the Brain* is broad and ambitious, with coverage on topics ranging from the number of ATP molecules consumed per bit in a blowfly photoreceptor (Chapter 3) to the logarithmic relationship between electromagnetic brain responses and auditory event probabilities in humans (Chapter 13). However, rather than being just another member of the ever-proliferating class of conference proceedings masquerading as books, this edited volume manages

to present a relatively integrated picture of current research in the field.

The introductory chapter (*Chapter 1*) by Baddeley lays the groundwork for the book by reviewing basic ideas and definitions in information theory. The use of down-to-earth examples provides ample intuition for concepts such as entropy and information. There is a useful review of practical methods for calculating entropy and mutual information, along with a succinct summary of information maximization, a topic that has received considerable attention in recent years in the context of sensory coding. Baddeley concludes with an objective enumeration of potential problems with information theory as an investigative tool, including the need for vast amounts of data and the distinction between information based on statistics and information that is of interest to the animal.

The rest of the book is divided into four parts, addressing four major themes of research in applying information theory to neurobiology and modeling: biological networks, artificial neural networks, psychological modeling, and formal analysis of biological systems. Such a division could very well be a reflection of corresponding sessions in the aforementioned conference but in the introductions that are provided for each part, we once again see evidence of the editors' conscious efforts to make the book more than a conference proceedings. In these introductions, the reader is provided with succinct summaries of the chapters as well as a useful glossary of technical terms and concepts that are central to the chapters in each part.

Part One (Biological Networks) focuses on how information theory can used to understand neural coding in the early stages of vision. The emphasis on early vision is not surprising given that recent efforts to explain retinal receptive field properties using information theory have met with considerable success. *Chapter 2* by Burton provides an excellent review of much of this research. The topics covered include the seminal work of Laughlin, van Hateren, and others on redundancy reduction and noise reduction in the fly retina and recent theories of retinal center-surround receptive fields in terms of collective coding and predictive coding. There is also a thought-provoking discussion of how the work of Atick and colleagues on retinal decorrelation [2] can be reconciled with Field's arguments for sparse coding based on the higher-order statistics of natural images [3]. The importance of higher-order statistics is discussed in greater detail in *Chapter 4* by Thompson who shows that an analysis based on the third-order correlation function of natural images leads to a better description of the distribution of spatial-frequency bandwidths in primate visual cortex than a second-order (power spectrum) based analysis.

Sandwiched between these two chapters on early vision and natural image statistics is *Chapter 3* by Laughlin and colleagues on the metabolic costs of information transmission and coding in

the blowfly retina. The molecular level of analysis adopted by this chapter makes it seem a bit out of place but the authors succeed in painting a harmonious picture of the marriage between information theory and bioenergetics. Buttressed by a variety of assumptions and approximations, the analyses lead to quantitative estimates of information transmission rates and ATP costs per bit for photoreceptors, interneurons, synapses, and action potentials in the blowfly retina. An especially interesting result is that action potentials are as costly as graded (analog) potentials. A reason for its use over short distances then could perhaps be the suppression of synaptic noise in convergent circuits. Although these results are encouraging, the problems involved in scaling up the ATP-based approach to higher-level visual processing and behavior appear to be considerable, but maybe not insurmountable.

One area to which information theory has contributed immensely is the formulation of appropriate optimization functions for learning in artificial neural networks. This area and related topics are the subject of *Part Two* of the book. *Chapter 5* by Harpur and Prager provides an excellent overview of recent attempts to formalize Barlow's idea of redundancy reduction [4] in terms of minimizing the statistical dependencies between a network's outputs. Such an approach has lead to new techniques such as independent component analysis (ICA) [5] and has allowed better characterizations of neurobiological receptive field properties than traditional techniques such as principal component analysis (PCA) and pure Hebbian learning [6]. Harpur and Prager describe a linear unsupervised network (recurrent error correction (REC) network) derived from the dual goals of minimizing input reconstruction errors and encouraging sparse representations. The network is essentially identical to the one proposed by Olshausen and Field for learning localized receptive fields from natural images [6], but by deriving the network from an information theoretic perspective, Harpur and Prager are able to show directly how encouraging sparseness also helps to reduce the mutual information between outputs, leading to greater independence between them. Their results from synthetic data, image coding, and speech coding are promising, although it remains to be seen if such an approach can be extended to the more common case of non-linear generative models for inputs.

A new perspective on the genesis and functional relevance of ocular dominance and orientation maps in the visual cortex is offered by Luttrell in *Chapter 6*. Starting from a probabilistic topology-dependent objective function, Luttrell derives a "soft encoder" network similar to Kohonen's self-organizing map (SOM) [7] (see also [8; 9; 10]). For both synthetic data and a Brodatz texture image, Luttrell's network develops cortex-like ocular dominance maps. For the texture image, an orientation map also emerges. However, the oriented receptive fields are not as localized

as those obtained by Harpur and Prager in Chapter 5, which leaves open the question of how these two approaches can be reconciled. *Chapter 7*, by Mato and Parga, describes a two-layer network model of cortical receptive field dynamics. The horizontal connections between the output units are adapted using an ICA-based learning algorithm proposed by Hérault and Jutten [11] for removing second- and higher-order correlations in the inputs. Mato and Parga show that their network can explain several physiological and psychophysical results such as the expansion of cortical receptive fields after adaptation to inputs with an artificial scotoma and perceptual shifts in feature localization tasks after such adaptation. Curiously, a discussion of this chapter is missing in the editors' introduction to Part Two.

Chapters 8 and *10*, by Wallis and Elliffe respectively, focus on a model of invariant object recognition. The cornerstone of the model is a Hebbian learning rule that is based on the trace of recent activity and that allows different views of an object to be associated based on temporal correlations. Although not discussed by the authors, the trace of activity can be regarded as implementing a type of statistical filter (e.g., the Kalman filter [12]) and therefore, a more general form of the trace rule could perhaps be derived by starting from a statistical viewpoint. *Chapter 9* by Krüger et al. also focuses on object recognition. They investigate a model based on curvature-sensitive generalized Gabor wavelets, nicknamed "banana wavelets" due to the resemblance to their tropical namesake. Although the link to information theory is a bit tenuous in this chapter, several interesting observations are made, especially regarding the need for features more complex than the localized Gabor wavelets obtained from sparse coding/ICA networks and the possibility of keeping the representation sparse at higher representational levels to aid in the search for relationships between object features.

Part Three contains three chapters on applications in psychology. *Chapter 11* by Aylett examines a mixture-of-Gaussians model for clearly articulated speech that can be used to analyze variations in clarity during normal spontaneous speech. In *Chapter 12*, Bullinaria reviews connectionist models of psychological phenomena such as developmental bursts, frequency effects, and speed-accuracy tradeoffs, and cautions that many such results are "free gifts" of connectionist networks i.e., they can be obtained with virtually any type of neural network model. *Chapter 13* by Sinkkonen derives a quantitative "law" specifying that resources in a discrete stationary environment should be allocated according to a logarithmic function of the probability of a stimulus. Although the law might not be entirely surprising given its similarity to Shannon's classical definition of information [13], the chapter provides a compelling experimental demonstration of its plausibility by showing

a logarithmic relationship between electromagnetic brain responses and auditory event probabilities in humans listening to unpredictable tone sequences. Among the chapters in Part Three, only chapter 13 focuses directly on the theme of the book, namely, the role of information theory in understanding the brain.

The final part of the book dwells on cortical, hippocampal, and neuronal modeling. *Chapter 14* by Schultz et al. describes an analytic model of information transmission from the area CA3 to area CA1 in the hippocampus (via the Schaffer collaterals). After several simplifying assumptions, their model predicts that information transmitted is maximal for binary firing rate distributions in CA3, and that the ratio between the numbers of neurons in CA1 and CA3 needs to be at or above 2 to allow CA1 to capture most of the information in CA3. The biological data on these predictions are however ambiguous, making it hard to judge the tenability of the model. This model is extended in *Chapter 15* by Mari et al. to include the entorhinal cortex (EC) and the perforant input pathway to CA3 and CA1. The mutual information between inputs in EC and outputs in CA1 is shown to increase approximately linearly with the number of perforant pathway fibers as well as with the fraction of the input cue in EC. Once again, the lack of comparisons with biological data makes its hard to evaluate the model. The counterintuitive phenomenon of stochastic resonance is the subject of *Chapter 16* by Bressloff and Roper. They show that for weak periodic inputs, the information transfer rate in a binary-threshold neuron with positive feedback can be maximized using a noisy threshold. The final chapter (*Chapter 17*) by Plumbley recasts within an information theoretic framework the old problem of magnification factors in cortical maps i.e., why certain input regions (such as the fovea) are represented by a much larger area in the cortex than other regions. In this chapter, magnification factors are viewed as arising from an attempt to maintain a uniform "information density" in the cortex given a non-uniform information density in sensory input regions. Unfortunately, the inability to define a suitable information density measure prevents the author from deriving concrete experimental predictions from this hypothesis.

While the book is noteworthy for its breadth of coverage, the diversity of the topics also at times hinders the cohesiveness of the volume. As mentioned above, a few of the chapters make little or no direct use of information theory and seem a bit out of place in a book devoted specifically to "information theory and the brain." The ordering of some of the chapters is also puzzling; for example, Chapters 8 and 10 are both on a trace-based learning model but are separated by a chapter on recognition using banana wavelets. In addition, most chapters make few or no references to the other chapters, thereby missing a potential opportunity for making novel connections between the

various ideas described in the book. However, the introductions written by the editors to each part of the book compensates to a certain degree for this lost opportunity. In several instances, their comments on how the techniques and results from one chapter relate to those from another help to unify the contents of the book. A final cause for concern is that only two of the twenty eight contributing authors are from research laboratories outside Europe. How representative of the field then is the research described in the book? Fortunately, much of the cutting-edge research in this field is indeed being pursued in European laboratories and it is reassuring to find several well-known leaders in the field among the contributing authors.

In summary, this book provides an excellent overview of how information theory has been used in recent years to understand both biological and artificial neural networks. It should be of interest to researchers and students in neural networks, artificial intelligence, computational neuroscience, and psychology. Several of the chapters are especially ideal for beginners seeking to teach themselves the basics of information theory (e.g., the introductory chapter by Baddeley) and its successful application in areas such as sensory coding (e.g., Chapters 2 and 5). Given the progress described in this book and the increasing use of information theoretic methods in neuroscience, there can be little doubt that the field is vibrant and active today, despite its soporific origins in a sleeping cat.

References

[1] F. Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193, 1954.

[2] J. J. Atick. Could information theory provide an ecological theory of sensory processing. *Network*, 3:213–251, 1992.

[3] D. J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

[4] H. B. Barlow. Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith, editor, *Sensory Communication*, pages 217–234. Cambridge, MA: MIT Press, 1961.

[5] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13:411–430, 2000.

[6] 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.

[7] T. Kohonen. *Self-Organization and Associative Memory*. Berlin: Springer, 1984.

[8] S.-I. Amari. A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*, 16:299–307, 1967.

[9] C. von der Malsburg. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14:85–100, 1973.

[10] G.A. Carpenter and S. Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, 37:54–115, 1987.

[11] J. Hérault and C. Jutten. *Reseaux neuronaux et traitment du signal*. Paris: Hermes, 1994.

[12] R. E. Kalman. A new approach to linear filtering and prediction theory. *Trans. ASME J. Basic Eng.*, 82:35–45, 1960.

[13] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.