Inhibition and Brain Computation

Steven L. Small*† and Gerhard H. Fromm*

Cognitive Modelling Laboratory
Department of Neurology* and Program in Intelligent Systems†
University of Pittsburgh
325 Scaife Hall, Pittsburgh, PA 15261
sls+@cs.cmu.edu

Abstract

The synapse plays a fundamental role in the computations performed by the brain. The excitatory or inhibitory nature of a synapse represents a (simplified) characterization of both the synapse itself and the computational role it plays in the larger circuit. Much speculation concerns the functional importance of excitation and inhibition in the physiology of the cerebral cortex. The current study uses neural network (connectionist) models to ask whether or not the relative proportion of inhibition (i.e., inhibitory synapses) and excitation (i.e., excitatory synapses) in the brain affects the development of its neural networks? The results are affirmative: An artificial neural network, designed to perform a particular task involving winner-take-all output nodes, is sensitive to the initial configuration of positive (excitatory) and negative (inhibitory) connections (synapses), such that it learns considerably faster when started with 60-75% inhibitory connections than when it includes a greater or lesser proportion than this. Implications of this result for neuroanatomy and neurophysiology are discussed.

Introduction

The brain computes through a distributed network of discrete neural elements whose pattern of connections gives rise to particular types of computations. Many morphological and functional features of these objects contribute to their ability to compute, with the synapse playing a fundamental role [Shepherd and Koch, 1990]. The excitatory or inhibitory nature of a synapse represents a (simplified) characterization of both the synapse itself and the computational role it plays in the larger circuit. The functional importance of excitation and inhibition in the brain is the subject of significant speculation [Fromm. 1992], which has led to assertions about the importance of inhibition in the physiology of the brain, particularly within the cerebral cortex.

In this paper, we use a neural network (connectionist) model to examine the question: Does the relative proportion of inhibition (i.e., inhibitory synapses) and excitation (i.e., excitatory synapses) in the brain affect the computational efficiency of its neural networks? Two parallel issues devolve from this question, one involving the development of computational circuits, and the other concerned with the operation of already learned circuits.

We investigate these questions in the context of the cortical visual system, particularly the results of Mishkin and his colleagues [Mishkin, et al., 1983] on macaque visual processing. When required to perform the dual task of visual object recognition and spatial localization, the macaque uses two separate visual systems to perform the two tasks, a temporal "what" system and a parietal "where" system [Desimone, et al., 1985; Mishkin, et al., 1983]. Rueckl, Cave, and Kosslyn [1989] have constructed a computer model of this system, which was used as a testbed for the present study of inhibition and excitation.

Two hypotheses motivated the current study: (1) The development of the visual system (to perform the object recognition and spatial localization task) takes place faster when the initial neural network contains a predominance of inhibitory synapses; and (2) Fully developed neural networks of the (two pathways of the) visual system operate more accurately when containing predominantly inhibitory synapses. We tested these two hypotheses by teaching numerous initial configurations of the Rueckl model (with different fractions of excitatory and inhibitory synapses) to perform the visual task.

This model represents one of a class of neural network (or connectionist) models currently under investigation by researchers from diverse disciplines. Such models attempt formally to understand biological neuronal networks, at the levels of both individual neuronal processing (e.g., dendritic computations, synaptic behavior) and of large assemblies of neuronal processing (e.g., cerebellar cortex) [Sejnowski, et al.,

Acknowledgment: The support of the NIH-NIDCD under grant number DC00054-02 is gratefully acknowledged.

1988; Sun, et al., 1988]. In addition, neural network models of high level cognitive processing, in such areas as vision [Feldman, 1989] and language [McClelland and Rumelhart, 1986; Seidenberg and McClelland, 1989], are reshaping accepted notions in information processing psychology. While these models differ greatly in the formal specifications of their neuronal units, and in the particular manner in which the units are connected to perform computations, they share similar underlying principles of organization.

The overall goal of the experimental method is to use a theoretical analysis to make suggestions for empirical scientists. As a necessity of the approach [Churchland and Sejnowski, 1987], we address computational questions about the brain at a high level of abstraction. Thus, we will not be able to show how much inhibition is actually used in a particular area of the brain or for a particular neurological task. Nonetheless, we do provide some suggestions about how the overall balance of inhibitory and excitatory synapses might make a difference in the computations that are possible. While this will not answer the morphological questions, it may help motivate research to establish a better correlation between anatomical and physiological results.

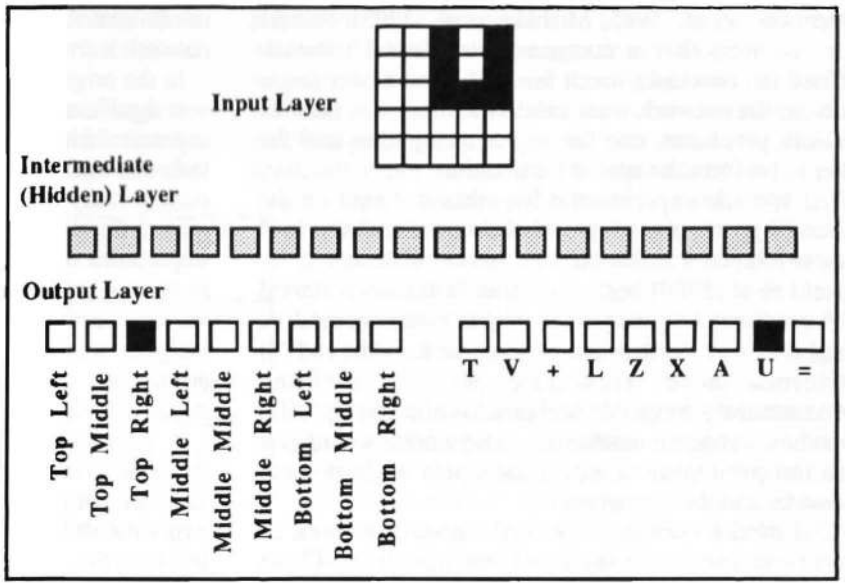

Figure 1: Network Representation of Visual Pattern Recognition

Methods

As with most connectionist models, and with all models at this level of analysis, many simplifying assumptions are made about neurons and synapses [Sejnowski, et al., 1988]. The model presented here uses a very simple model of a neuron, and an even simpler model of a synapse. These simple models are shown in Table 1, and are representative of the strategies common in connectionist modelling. (See our previous discussion of this in [Small, 1991].) A parallel distributed processing (PDP) approach [Rumelhart and McClelland, 1986], employing a layered feed forward (i.e., non-recurrent) network with one hidden layer of computational units in addition to the input and output layers, formed the basis of the implementation.

Such networks are able to learn by example to perform some task (i.e., associating a number of input patterns with desired output patterns) by adjusting their connection strengths according to particular error minimization rules. The back propagation learning algorithm [Rumelhart, et al., 1985] represents a useful learning strategy. One criterion for the success of a PDP model using back propagation is the number of presentations of each training example required to teach the model to perform the desired task. The models discussed in this paper use this criterion; a network that learns a task quicker (i.e., with fewer presentations of each training instance) is considered to be superior to one that learns the task slower. Of course, these numbers are subject to statistical interpretation, and the concepts of faster and slower must conform to standard criteria of significance.

As noted, the experiments presented here were conducted with the visual system model of Rueckl and his colleagues [1989]. The model performs the classification of two-dimensional visual images into two categories, one representing what object was shown and the other representing where on the input grid the image appeared.

CNS Concept	Model Analogue	Nature	Description
Neuron	Unit	Abstraction	Associated values and functions
Synaptic strength	Connection weight	Value	Real number
Axon firing rate	Unit potential	Value	Real number
Synapse	Unit input	Value	Weighted unit potential
Inhibition	Negative weight	Value	Negative real number
Excitation	Positive weight	Value	Positive real number
Depolarization	Potential function	Function	Adjusted sum of inputs
Threshold	Bias	Value	Real number

Table 1: Computer Model Correlates of Neurobiological Concepts

This dual task of visual object recognition and spatial localization has a neurobiological basis in the temporal "what" system and parietal "where" systems identified in the macaque monkey

[Desimone, et al., 1985; Mishkin, et al., 1983]. Rueckl et al showed that a computational neural network learned the two tasks much faster if instead of a single process, the network were subdivided into two parallel network processes, one for object recognition and the other to perform the spatial localization.

The specific experimental hypotheses, based on the general hypotheses discussed above, are that the feed forward layered connectionist network designed by Rueckl et al [1989] both (1) learns faster when started with predominantly negative weights than with random weights or with predominantly positive weights; and (2) performs more accurately when containing predominantly negative weights. While the specific hypotheses concern mathematical networks, we suggest that the principles of operation apply to biological networks and their synapses.

This model consists of a feed forward network of units containing three layers: (1) an input layer, (2) an output layer, and (3) an intermediate (hidden) layer. These three layers are illustrated graphically in Figure 1. The input layer consists of a linear representation of a two dimensional visual pattern. The 5 x 5 input grid of Figure 1, representing the letter "U", is actually represented as the linear vector of binary digits, with a "1" representing a pixel in the pattern and a "0" one that is not in the pattern. The output layer represents *what* pattern was presented and *where* in the two dimensional input space the pattern was presented.

The output layer of Figure 1 illustrates these two sublayers of representation: One sublayer (the right hand part of the output layer in the Figure) contains the information on what input object was presented — in this case the letter "U" — and the other sublayer (the left hand part of the output layer in the Figure) contains the information as to where in the input grid that object appeared — in this case, in the top right position of the input grid. The network representation of the output layer uses binary digits, with a "1" for the correct identification and location, and a "0" otherwise.

The units of one layer are fully connected to those of the next layer, and the connections can be either positive (excitatory) or negative (inhibitory). The units within a particular layer are not connected. The network starts with random connection strengths among the units of the adjacent layers. It is then repeatedly presented with nine different input patterns in all nine possible positions, along with the desired output values, indicating what pattern was presented where. The example input pattern and correct output value of Figure 1 are illustrative. The model then uses the back propagation algorithm [Rumelhart et al, 1985] to change the connection strengths (weights) of the network in a way that minimizes the overall network error. Ultimately the network learns to classify all the input patterns.

In the original model of Rueckl et al [1989], learning was significantly faster when the network was split into separate "what" and "where" systems than when the task was attempted by a single undivided network. The current study used the split network for all trials.

In order to test the hypothesis about the relative importance of inhibitory versus excitatory connections in brain computations, another network parameter was varied, namely, the percentage of initial network weights with positive values. Recall that positive connection weights represent excitatory synapses, and negative weights represent inhibitory synapses.

A pseudo-random number generator was used to generate two values, a real number between 0 and +2, and an integer sign (either -1 or +1). Thirteen experiments were conducted: For each connection in the network, the probability of it receiving an initial positive weight was 0% in one trial, 6.25% the next trial, and 12.5%, 18.75%, 25%, 31.25%, 37.5%, 43.75%, 50%, 56.25%, 62.5%, 68.75%, and 75% in the twelve additional trials. The learning algorithm was constructed to present input/output pairs (training instances) repeatedly until either (a) the sum squared error of the network dipped below 4.0; or (b) the total number of presentations of the entire corpus of training instances (one epoch = 9 images x 9 positions = 81 individual training instances) reached 200. These numbers were chosen following several pilot experiments that showed that a network error of about 4.0 represented good performance. The limit to 200 epochs was a practical decision motivated by limitations in computational resources.

For this project, the basic model was reimplemented using the DYSNET simulator. Specific choices regarding potential functions, learning parameters, error measure, and weight updating function are shown in Table 2. Note that these choices may or may not reflect those of the original model by Rueckl et al [1989] and are practically, but not theoretically, important.

Results

The results of these experiments are summarized in

Attribute	Value	Reference
Network Structure	Feed Forward	[Rumelhart and McClelland, 1986]
Hidden Layers	One	[Rumelhart and McClelland, 1986]
Layer Widths	25 x 18 x 18 units	[Rueckl, et al., 1989]
Substructures	Splitting	[Rueckl, et al., 1989]
Potential Function	Logistic Function	[Rumelhart, et al., 1985]
Learning Algorithm	Generalized Delta Rule	[Rumelhart, et al., 1985]
Weight Updating	QuickProp Algorithm	[Fahlman, 1988]
Error Measure	Sum Squared Error	[Rumelhart, et al., 1985]
Learning Rate	0.5 ÷ Unit Fan In	[Fahlman, 1988]

Table 2: Computational Features of the Model

Table 3. These data reflect the average of one hundred individual learning experiments at each fraction of initial positive weights.

Note that both the average network error and the average number of epochs vary with the fraction of positive initial random weights, reaching a nadir between 25% and 37.5%, but increasing as the fraction of initial positive weights decreases below or increases above this level. The statistical results compare the number of epochs required to learn the task at each starting configuration (i.e., percentage of initial positive weights) with the minimum number required when 25% of the initial weights are positive. The Student t-test using a two tailed distribution was used for this comparison. When adjusted for multiple comparisons, it still shows a significant effect: A starting configuration reflecting a preponderance of negative weights (within a specific range) leads to faster network convergence than with a preponderance of positive weights.

Fraction Positive Weights	Mean Network Error	Mean Number Epochs	Probability (number of epochs)	Paired T Value
0.125	5.290	149.51	5.649	$p < 0.0001$
0.250	4.183	111.93	***	***
0.375	4.095	117.49	0.869	$p = 0.3824$
0.500	4.276	150.87	6.236	$p < 0.0001$
0.625	5.128	181.10	13.129	$p < 0.0001$
0.750	7.181	197.18	17.977	$p < 0.0001$

Table 3: Experimental Results

Figure 2 shows a graphic illustration of one portion of the initial configuration when the fraction of initial positive weights was set to 25% of all connection weights. In the Figure, white squares represent positive values and black squares represent negative values. The area of the box represents the real number value (in this example, the largest box encodes an absolute value of 2.0). The Figure illustrates the connection strengths between each of the eighteen hidden units and the fifth unit of the output layer.

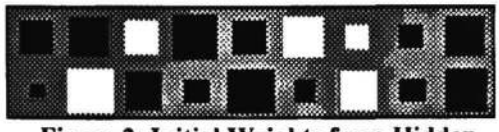

Figure 2: Initial Weights from Hidden Layer to Fifth Unit of Output Layer

Figures 3 and 4 illustrate graphically the results of the thirteen experiments. (Note that the standard errors of the means, which are not shown in the Figures, are extremely small). Figure 3 shows the average minimum network error achieved in 100 separate learning trials, when the starting configuration included random connection strengths in which the percentage of positive and negative values was varied. The abscissa of this graph measures the fraction of positive initial weights and the ordinate measures the sum squared error of the network. Note that the average network error reaches a minimum with a starting configuration of 37.5% excitatory weights, and increasing or decreasing this percentage sharply increases the total error (0% excitation is not shown: the error exceeds 8).

Figure 4 shows the average number of trials required to reach either a network error of 4.0 or a total of 200 trials. Numbers close to 200 therefore represent failure to converge in 200 trials. As in Figure 3, these data were accumulated from 100 separate learning trials, with a varied initial percentage of positive and negative connection weights. The abscissa of this graph measures the fraction of positive initial weights and the ordinate measures the number of trials. The minimum number of epochs required to learn the task occurs at a starting configuration of 31.25% inhibition, with alterations in this percentage significantly impairing learning. Analysis of the final network demonstrated a linear correlation between the inhibition fraction of the initial network (before learning) and that of the completely trained network.

Discussion

The present study demonstrates that an artificial neural network, designed to perform a particular task, is sensitive to the initial configuration of positive (excitatory) and negative (inhibitory) connections (synapses). The particular network examined uses winner-take-all output representations, and learning is considerably faster when the structure of the network includes 60 - 75% inhibitory connections than when it includes a greater or lesser proportion than this. While there are many intuitive analyses of the importance of inhibition for brain computations, both at the level of individual neurons as well as at the level of the

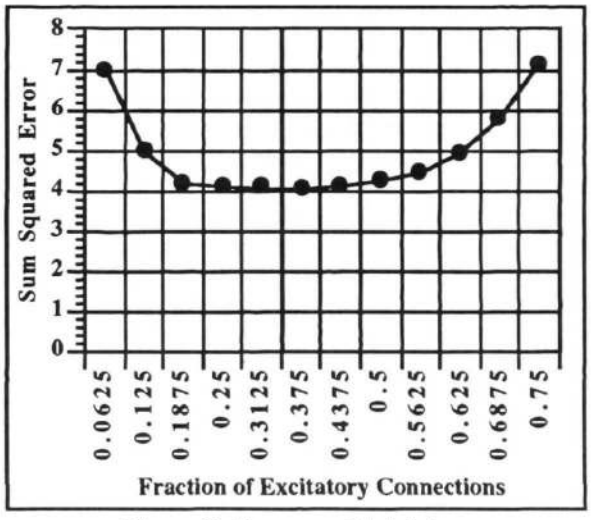

Figure 3: Error vs. Weights

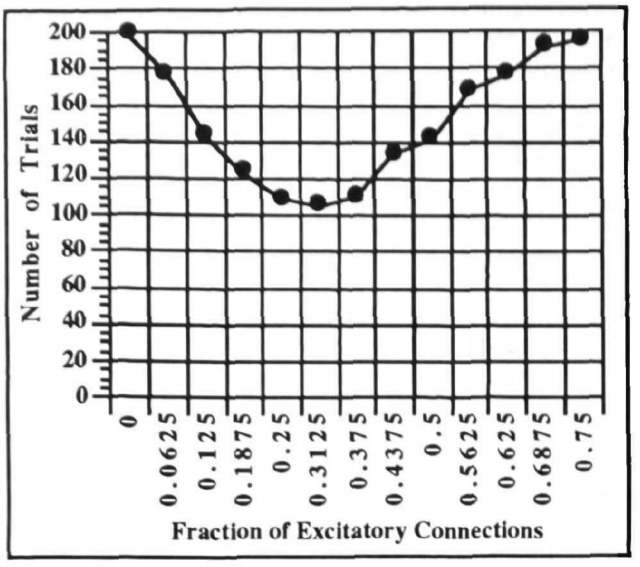

Figure 4: Trials to Converge vs. Weights

organism as a whole, there has not been a similar suggestion from computational simulation.

A complex set of events must occur in a neuron for it to initiate an electrical signal, and this depends largely on the architecture of the individual neuron and the nature of the chemical signals it receives. Each afferent signal (neuronal input) can be viewed as having either excitatory (facilitatory) or inhibitory effects on the development of an action potential. The signals constituting these inputs manifest a variety of interesting temporal and spatial organizations as well as complex local interactions [Koch, et al., 1983], all of which contribute to their ultimate computational conclusion — whether or not to initiate an action potential.

While the relative proportion of inhibitory and excitatory synapses in the central nervous system is not known, attempts have been made to quantify these proportions by a variety of methods. Immunocytochemical analyses have led to the view that gamma-amino butyric acid (GABA) is the most prevalent inhibitory neurotransmitter of the central nervous system [Kandel and Schwartz, 1985]. Smith [1989] even suggests that GABA is the most widely used neurotransmitter of any kind in the CNS, with over 40% of all synapses using GABA.

Support for this idea comes from studies of cortical interneurons, which suggest that most are GABAergic. These studies, summarized by Jones and Hendry [1986], use three different techniques in arriving at the conclusion: (a) [^{3}H]GABA uptake; (b) immunoreactivity for GABA; or (c) immunoreactivity for glutamic acid decarboxylase (GAD).

In the prestriate visual system of the macaque, the lateral geniculate nucleus contains significant immunoreactivity for glutamic acid decarboxylase (GAD), an enzyme required for GABA synthesis [Shaw and Cynader, 1986]. In the optic tectum of the frog, nearly one third all tectal cells are immunoreactive for GABA. In the striate cortex of the macaque monkey, layers 2, 3, 4A, and 4C contain large concentrations of GABA receptors [Shaw and Cynader, 1986].

Physiological study has led to the notion of orientation selectivity as a fundamental organizing principle of the visual cortex [Hubel and Wiesel, 1962]. The computational implementation of orientation selectivity requires that a bar of excitation be surrounded by a massive ring of inhibition, in order to eliminate ambiguity in the perception of an edge in the desired orientation. This computational constraint suggests that a large number of synapses act principally in an inhibitory manner.

Pharmacological evidence to support this postulate comes from studies of the selective GABA antagonist bicuculline. Application of bicuculline to orientation selective nerve cells in the cortex of the cat abolishes their response to the correctly oriented bar of light [Sillito, 1986].

Investigation of the ultrastructural (anatomical) differences between two types of synapses has led to different results. Type I (round asymmetric or RA) synapses, which are frequently excitatory, have asymmetrical densification of their pre- and post-synaptic membranes, and are associated with round synaptic vesicles. Type II (flat symmetric or FS) synapses, frequently inhibitory, have symmetrical densification and have flattened or pleomorphic vesicles [Gray, 1959; Shepherd and Koch, 1990].

Beaulieu and Colonnier [1985] studied the cat's visual cortex using these methods and concluded that about 84% of the synapses are of the RA type (usually excitatory) and 16% of the FS type (usually inhibitory). Two main questions remain in interpreting this data (and other data like it): (1) Do these RA synapses contain primarily an excitatory neurotransmitter, an inhibitory neurotransmitter, or both? (2) What is the relationship between the number of inhibitory synapses and the magnitude of their computational effects? Ultimately, we need to know the extent to which anatomical and physiological information bears on the computational issues and vice versa.

There is an apparent discrepancy between anatomical and physiological data regarding neuronal processing in the primary visual cortex of the cat. The anatomical results are particularly difficult to interpret, since knowledge of the number of synapses with a particular morphological structure does not necessarily indicate how these synapses are used computationally in the actual physiological setting. Combinations of excitatory and inhibitory synapses in complex topographical arrangements lead to intricate local circuit behaviors that may not correlate in any simple way with their absolute numbers. For example, a single inhibitory synapse, appropriately placed, can negate multiple excitatory stimuli.

Conclusion

While the integration of ultrastructural, physiological and computational data may require the development of new techniques, the goal of doing so may have important consequences for the understanding of structure/function relationships. Using computer modelling techniques and abstract representations of neurons and synapses, the present study suggests a preeminent role for inhibition in the computational organization of the brain.

In the brain, local circuit organization of inhibitory synapses, regardless of their absolute numbers, can have a controlling effect. When these are located closer to the soma than the excitatory synapses, they can (under certain circumstances) totally negate the excitatory effects [Koch, et al., 1983; Shepherd and Brayton, 1987; Shepherd and Koch, 1990]. Whether or not this computational effect bears on the situation in the visual cortex is not clear. However, the computational and physiological data suggest that the apparent preponderance of (typically excitatory) RA synapses in this area does not correlate with a preponderance of overall excitatory activity there.

The present study was initiated in response to speculation about the importance of inhibition in the physiological function of the human brain [Fromm, 1992]. The modelling results demonstrate a highly significant role of inhibition in particular artificial neural networks (containing sparsely coded output representations) and support the concept of inhibition as a basic computational feature of the brain.

References

Beaulieu, C. and M. Colonnier (1985): A Laminar Analysis of the Number of Round-Asymmetrical and Flat-Symmetrical Synapses on Spines, Dendritic Trunks, and Cell Bodies in Area 17 of the Cat, *J Comp Neurol*, 231:180-189.

Churchland, P. S. and T. J. Sejnowski (1987): Neural Representation and Neural Computation, The Johns Hopkins University.

Desimone, R., S. J. Schein, J. Moran and L. G. Ungerleider (1985): Contour, Color, and Shape Analysis Beyond the Striate Cortex, *Vision Res*, 25:441-452.

Fahlman, S. E. (1988): An Empirical Study of Learning Speed in Back-Propagation Networks, Carnegie Mellon University.

Feldman, J. A. (1989): Neural Representation and Neural Computation, in *Neural Connections, Mental Computation*.

Fromm, G. H. (1992): Neurophysiological Speculations on Zen Enlightenment, *J Mind Behav*, 13(2):163-168.

Gray, E. G. (1959): Axo-somatic and Axo-dendritic Synapses of the Cerebral Cortex: An Electron Microscope Study, *J Anat*, 93:420-433.

Hubel, D. H. and T. N. Wiesel (1962): Receptive Fields, Binocular Interaction, and Functional Architecture of Monkey Striate Cortex, *J Physiol (London)*, 160:106-154.

Jones, E. G. and S. H. Hendry (1986): Co-Localization of GABA and Neuropeptides in Neocortical Neurons, *TINS*, 9:71-76.

Kandel, E. R. and J. H. Schwartz (ed.) (1985): *Principles of Neural Science (Second Edition)*.

Koch, C., T. Poggio and V. Torre (1983): Nonlinear Interactions in a Dendritic Tree: Localization, Timing, and Role in Information Processing, *Proc Natl Acad Sci USA*, 80:2799-2802.

McClelland, J. L. and D. E. Rumelhart (1986): *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Volume 2: Psychological and Biological Models.*

Mishkin, M., L. G. Ungerleider and K. Macko A. (1983): Object Vision and Spatial Vision: Two Cortical Pathways, *Trend Neurosci*, 6:414-417.

Rueckl, J. G., K. R. Cave and S. M. Kosslyn (1989): Why are "What" and "Where" Processed by Separate Cortical Visual Systems? A Computational Investigation, *J Cog Neurosci*, 1:171-186.

Rumelhart, D. E., G. E. Hinton and R. J. Williams (1985): Learning Internal Representations by Error Propagation, University of California San Diego.

Rumelhart, D. E. and J. L. McClelland (1986): *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Volume 1: Foundations.*

Seidenberg, M. S. and J. L. McClelland (1989): A Distributed, Developmental Model of Word Recognition and Naming, *Psych Rev*, 96:523-568.

Sejnowski, T., C. Koch and P. Churchland (1988): Computational Neuroscience, *Science*, 241:1299-1306.

Shaw, C. and M. Cynader (1986): Laminar Distribution of Receptors in Monkey (Macaca Fascicularis) Geniculostriate System, *J Comp Neurol*, 248:301-312.

Shepherd, G. M. and R. K. Brayton (1987): Logic Operations are Properties of Computer-Simulated Interactions between Excitable Dendritic Spines, *Neuroscience*, 21:151-166.

Shepherd, G. M. and C. Koch (1990): Introduction to Synaptic Circuits, in *Synaptic Organization of the Brain*.

Sillito, A. M., Functional Considerations of the Operation of GABAergic Inhibitory Processes in the Visual Cortex, in *Cerebral Cortex, Volume 2: Functional Properties of Cortical Cells.*

Small, S. L. (1991): Focal and Diffuse Lesions of Cognitive Models, *Proceedings of the Thirteenth Annual Meeting of the Cognitive Science Society.*

Smith, C. U. M. (1989): *Elements of Molecular Biology.*

Sun, R., E. Marder and D. Waltz (1988): The Modelling of Lobster Stomatogastric Neural Networks, Brandeis University.