Focal and Diffuse Lesions of Cognitive Models

Steven L. Small

Cognitive Modelling Laboratory
Department of Neurology
University of Pittsburgh

Abstract

With the recent ability to construct fault tolerant computer models using connectionist approaches, researchers are now able to investigate the effects of damage to these models. This has great appeal for cognitive science as it provides a further way to verify or falsify a computer model. Existing studies employ a concept of network "lesioning" that fails to have explanatory adequacy for neurobiology. While using anatomically plausible architectures for cognitive models, they nonetheless use biologically implausible methods for simulating neurological damage to these networks. This paper examines the different objects of computational networks and their analogical neurobiological counterparts, and suggests a taxonomy of connectionist network lesion methods. Finally, an existing visual system model is used as a testbed to study the differential effects of focal and diffuse lesions. The experiments with focal damage versus diffuse damage suggest that while the effects of focal brain injury may be due to the particular computations performed in some brain area, the effects of diffuse brain injury or degeneration may cause cognitive deficits because of the inherent nature of the brain as a distributed computational device, and not through differential local effects.

Part I: Introduction

One feature of connectionist networks that makes them particularly interesting for cognitive modelling is their computational relationship to the human brain. This relationship, expressed to a greater or lesser degree in different types of network architectures and domains of modelling (i.e, cognitive and/or neurobiological), contributes a new class of constraints to the validation of such models (Sejnowski and Churchland, 1988). One valuable new constraint is that of lesionability. Instead of evaluating a model solely on its performance with respect to the normal processing behavior of the object being modelled (e.g., input/output behavior, intermediate representations), a model may now be subject to computational disruption and be expected to produce behaviors that are analogous to known abnormal processing behaviors as well.

The analogy between the computational structures of a cognitive model and the neurobiological structures of nature can be of variable strength, due in part to (a) the nature of the computational architectures themselves; and (b) the specification of the analogical relationships by the modeller. This holds equally for the lesions inflicted on computational networks in the name of neurological or cognitive impairment. In this paper, we explore the ways in which a connectionist network can be lesioned, and discuss their plausibility in terms of both basic neurobiology and clinical neurology.

Address: Cognitive Modelling Laboratory, Department of Neurology, University of Pittsburgh, 325 Scaife Hall, Pittsburgh, PA 15261. Phone: (412) 648-9200. Network: sls@dsl.pitt.edu.

Neurological Lesions

Acquired damage to the human brain leads to dysfunctional performance in a variety of modalities, depending on both the quality of the damage and its quantity. The damaged area of the brain is considered the "lesion" (Damasio and Damasio, 1989) and neurologists focus on using various intellectual and radiographic techniques to characterize the location, size, and cause of a particular lesion. In the case of visual or linguistic disorders, neuropsychologists employ specialized examination techniques to make these inductions (Boller and Grafman, 1990).

While it has always been questioned to what extent different functions localize to different parts of the brain, the currently prevailing view is that there is in fact a tremendous localization of function (Galaburda, et al., 1978). Such a view is necessary for the inductive reasoning steps described above to be meaningful. Although localization of function definitely exists to some extent (e.g., primary motor area), there is ample reason to ask two things: (1) What functions are the ones that localize? (2) Over how large an area do functions localize, given the probable distributed representations of the neural implementations of these functions? A corollary to question (2) is how minutely can we attempt to localize particular functions?

Many causes exist for brain lesions. Examples include strokes, tumors, intoxications, and degenerative diseases. A useful way to classify them is to divide them into two main categories: Focal lesions represent damage to well circumscribed regions of brain substance; while the extent of such lesions cannot always be perfectly demarcated, and frequently changes over time, there is nonetheless a focus of impaired nervous system functioning. Strokes constitute the most prevalent cause of focal brain lesions. Diffuse lesions involve damage to a large number of discrete neural elements over a widespread area of the brain, involving one or more particular classes of neurons or neuronal elements. Alzheimer's Disease is the most prevalent cause of diffuse brain lesions.

Study	Function	Damage	Nature of Model Disruption(s)
(a)	Dyslexia	Focal	(1) Proportion of weights set to zero (2) Random number added to all weights (3) Removal of some hidden units
(b)	Spatial neglect	Focal	Unspecified "damage" to "connections"
(c)	Schizophrenia	Diffuse	Decrease gain in particular subnetwork
(d)	Aphasia	Focal	Unspecified "noise" to "connections"

Table 1: Existing Computer Model Lesioning Experiments

Computational Model Lesions

With the recent ability to construct fault tolerant computer models using connectionist approaches, researchers are now able to investigate the effects of damage to these models. Several investigators have taken this approach to the study of cognition, and have lesioned particular cognitive models, and then attempted to compare the resulting model performance with that of people who have suffered brain lesions. This method has great appeal for cognitive science as it provides a further way to support or falsify a computer model.

This method has been attempted in several areas, including (a) acquired dyslexia (alexia) (Hinton and Shallice, 1989); (b) "neglect dyslexia" (spatial neglect in reading) (Mozer and Behrmann, 1989); (c) schizophrenia (Cohen and Servan-Schreiber, 1989); and (d) aphasia (dysphasia) (Miikkulainen, 1990). In each case, an interesting cognitive model was disrupted to produce one or more behaviors that resemble human information processing under some condition of neurological damage. Table 1 shows for each model whether the human condition reflects diffuse or focal damage and how the model was disrupted to simulate that condition.

Note that these studies employ a concept of network "lesioning" that fails to have explanatory adequacy (Chomsky, 1965) in the neurobiological sense. Whereas these researchers go to great trouble to build anatomically plausible architectures for cognitive models, they nonetheless use biologically implausible methods for simulating neurological damage to these networks. For example, while aphasia is typically the result of focal neurological damage (e.g., a stroke), Miikkulainen (1990) adds "noise" to the connections of his network, which is a diffuse strategy. Spatial neglect also arises most typically from focal brain damage, yet Mozer and Behrmann (1989) nonspecifically "damage" some connections. In applying their notion of "gain" in a diffuse manner to a subnetwork of their model, Cohen and Servan-Schreiber (1989) do in fact meet the explanatory criteria suggested here, as they are using diffuse lesions to account for impaired problem-solving behavior in patients with diffusely damaged brains (i.e., schizophrenia).

The study of Hinton and Shallice (1989) investigates the effects of three different methods of network lesioning. They applied each method to the different layers of their model to observe the effects, and explored empirically the relationships between the reading behaviors of differently impaired networks and dyslexic patients. While acquired dyslexia typically results from focal brain insults, the study explored diffuse model lesions (i.e., resetting a proportion of weights, adding a random number to all weights) in addition to a method of focal lesioning (i.e., removing some hidden units, which is only focal if these units are physically adjacent to each other).

Part II: A Taxonomy of Lesions

Connectionist models of cognitive processing typically incorporate architectures that limit specifically the types of lesions one might consider for computational experiments of dysfunctional cognitive performance. Table 2 lists some neurobiological concepts and their analogues in connectionist models. It does not matter that a single unit in a parallel distributed (PDP) cognitive model does not represent a single neuron in the brain (Sejnowski, et al., 1988). Biological nervous systems have motivated massively parallel approaches (Feldman, 1989) and in this paper, the

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 2: Computer Model Correlates of Neurobiological Concepts

analogy between neurobiological and computational networks will be analyzed in terms of lesions that can be produced.

For each biological entity shown in the Table, a variety of neurological disruptions naturally occur in human illness. Furthermore, in vitro and in vivo studies of the neuroanatomy, neurophysiology, and neuropharmacology of these entities have led to some valuable information.

Structures to Lesion
Processing Units
ConnectionWeights
Activation Functions

Table 3: Objects to Lesion

In focal neurological disease, whole collections of neurons, supporting structures, and connections are lost in one brain area. By contrast, in diffuse disease, one particular aspect of (one type of) neuronal functioning might be lost throughout the entire brain.

Lesion Objects

Lesions to network models can occur either focally or diffusely to the three general structures listed in Table 3. Each network object has a neurobiological correlate, and different lesions to these objects have neurological analogues.

The specific subpopulation for focal lesions depends on the architecture of the particular network representations in the model. One method of classification is by location in the overall network. Table 4 shows that the concept of location has several possible interpretations, all of which are interesting and relevant.

Locations
Functional Location in Cognitive System
Functional Location in Computational Processing
Spatial Location

Table 4: Object Selection by Location

The input and output layers of a network are locations from the vantage point of computational function. The subpart of the input layer that represents the phonological input lexicon for a model of lexical access is a location in the cognitive domain. The bottom left quadrant of a two dimensional drawing of a network is a topographically specified location. Note that this latter specification, which sounds more arbitrary and untheoretical, may in fact be the most accurate way of (focally) lesioning a cognitive (neuroscientific) model. This is a consequence of the vascular organization of the brain, which leads to brain lesions that follow a spatial pattern, rather than a functional one.

Values
Range of Absolute Values
Range of Signed Values
All Values of a Particular Sign

Table 5: Object Selection by Value

Dynamic aspects of computational networks include the values associated with the fixed structures over time. Table 5 lists several specific categories of values that can be used as selection criteria for choosing structures to lesion. Note that both ranges of values and their signs alone are useful.

Network Lesions
Ablation (Deletion)
Attenuation
Augmentation
Resetting
Addition of Noise

Table 6: Lesion Methods

The neurobiological plausibility of the different selections by value depends to some extent on the specific intended analogy between the lesioned structures and the brain or cognitive system. However, the existence of anatomical differences between (certain) inhibitory and excitatory synapses in the brain (Shepherd and Koch, 1990) suggests that a strategy based on signed values (or signs alone) may have validity for different classes of structures.

Lesion Methods

A network model can be computationally lesioned by a number of different operations. Table 6 lists a few possibilities, involving (a) removal of an object; (b) increasing or decreasing its value by a fixed percentage or by adding random noise; and (c) setting the value of an object to some previous value.

Units

Insofar as the computational units are analogous to neurons in the brain, they represent a well founded candidate population for lesioning experiments. Since a typical computational unit represents a collection of values and functions, the designation of a lesion must be more specific, e.g., unit output value, as shown in Table 7.

A straightforward disruption of the functioning of an individual unit is to delete it from the network. This is achieved by removing all of its connections, such that no further network processing includes computations originating there. Less drastic manipulations include increasing or decreasing its output by a certain percentage or a fixed amount or adding noise (random values within some range) to the output.

Unit deletion has a neurobiological correlate in destructive brain damage (e.g., stroke). The biological correlate of unit potential is the axonal firing rate. Augmentation and attenuation of this value are analogous to increases and

decreases in firing rate, which could be the result of changes in (a) neurotransmitter (or neuromodulator) concentration; (b) excitatory inputs; (c) inhibitory inputs; or (d) axonal conduction. Certain conditions lead to changes in neurotransmitter concentrations in particular brain areas, e.g., dopamine concentration in the striatum in Parkinson's disease. An analogy to random noise in affecting axonal firing rate might be brain intoxications of various kinds, either by external toxins (e.g., drugs) or internal ones (e.g., metabolic derangements).

Unit Lesion Sites
Potential (Output Value)
Input Values
Activation Functions

Table 7: Aspects of Units

The biological correlate of a unit input value is the synapse. Since the synapses of a neuron are spatially distributed along a variety of dendrites, it makes more sense to talk about subpopulations of units than either individual synapses or all synapses. Relevant subpopulations include the synapses on a particular dendrite or ones that utilize a particular neurotransmitter. Attenuation or augmentation of the analogous values in the model corresponds to the under- or over-sensitivity of particular synaptic transmission (e.g., post-synaptic receptor density), and could be caused by local neurotransmitter changes. Toxins could produce effects analogous to the addition of noise.

Activation Functions

Activation functions take unit inputs and produce an output. The neurobiological correlate of the activation function is membrane depolarization, and many different factors bear on the ability of the neuronal membrane to depolarize. While the absolute number of active synapses plays a role, perhaps more important are their spatial distribution and chemical characteristics. This was noted above, and lesions to subpopulations of unit inputs constitute reasonable analogies to neurological damage to neurotransmitter systems or dendritic locales.

Table 8 lists four aspects of activation functions that are subject to lesioning. Each of these aspects corresponds to a stage in the application of the function, and affects the ultimate behavior of the unit. The unit first computes the (weighted) sum of all inputs, corresponding roughly to the combined membrane electrical effects of the chemistry of the synapses. This number is then manipulated arithmetically to produce another value, which corresponds to the total membrane depolarization. In connectionist networks, the summed values are combined by a "squashing function" (Rumelhart and McClelland, 1986), and the precise sigmoidal curve obtained depends on a value called the "gain", the adjustment of which is the subject of the interesting lesioning experiments of Cohen and Servan-Shreiber (1989) discussed earlier. Lastly, action potential propagation depends on this final value exceeding a threshold, which can also be varied. The

Activation Function Lesion Sites
Unit Input Sum
Adjustment of Input Sum
Gain
Threshold (Bias)

Table 8: Activation Functions

presence of different physiological properties of neurons and responses to neuromodulators (Shepherd, 1990), e.g., peptides, provide ample neurobiological correlation of these mathematical manipulations.

Weights

The connection strengths or weights of an artificial neural network correspond to synaptic strengths of connections. Negative weights are analogous to inhibitory synapses, positive weights to excitatory synapses. Weight lesions thus correspond to disruptions in the role of particular synapses in effecting the action potential.

Part III: Experimental Results

We have recently been investigating network lesions in a general manner by incorporating lesioning into a connectionist simulator (called DYSNET) and using it to study various models in cognitive neuroscience. Lesion specification in DYSNET involves selection of one feature from each column of Table 9, and then declaration of the appropriate parameter(s). Particularly useful selection criteria include specific network partitions (subnetworks), which may be lesioned independently, and value ranges, either absolute (e.g., all weights with absolute value less than 2.0), signed (e.g., all values between -1.0 and 1.5), or signs themselves (e.g., all inhibitory weights).

Motivation

One computational experiment was motivated by an interesting difference in visual system performance under different conditions of damage. Focal damage to the parieto-occipital junction can lead to a syndrome of visuospatial

Object Type	Functional Location	Range	Lesion Type
Weight	Layer	Absolute Value	Deletion
Unit	Partition	Signed Value	Attenuation
		Sign	Reset
			Noise

Table 9: Parameters Characterizing Computational Lesions

disruption known as Balint's Syndrome (Balint, 1909). Patients with this syndrome are unable visually to guide their hands to a particular location in space in order to grasp an object. They are able to tell what object was shown to them, but seem to have difficulty in using visuospatial knowledge. Focal damage to the posterior temporal lobes can lead to object recognition problems, without visuospatial difficulties.

Patients with diffuse neurological damage from Alzheimer's Disease have problems with both object recognition and spatial orientation. However, problems with object recognition both precede the development of visuospatial dysfunction and are more serious than the spatial problems (Mendez, et al., 1990).

Two Pathways

In order to model this difference, a number of lesion experiments were conducted using the visual system model of Rueckl and his colleagues (Rueckl, et al., 1989). Their network classifies two-dimensional visual images into two categories, (1) what object was shown, and (2) where in the visual image the object appeared. The empirical studies of Mishkin and his colleagues (1983) on macaque visual processing constrained the architecture of the connectionist model and led to computational hypotheses. 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) showed that a computational neural network learned the two tasks much faster if the network were subdivided into two parallel networks, one to perform the object recognition and the other to perform the spatial localization.

Hypotheses

The conclusion of these researchers is that the learning of an object recognition and spatial localization task is easier with separate "what" and "where" networks than with a single integrated network. The current investigation of lesioning aims to build upon this research, and to test further its neurobiological plausibility by comparing its functioning when lesioned to several general features of impaired human functioning.

As noted, Balint's Syndrome (Balint, 1909) involves significant problems in visuospatial analysis (Newcombe and Ratcliff, 1989). Brain lesions that cause such problems are in the junction of the occipital and parietal lobes, and are typically the result of focal damage such as stroke. Patients with Alzheimer Disease (AD) can also get Balint's syndrome from their diffuse degenerative disease, but early deficits in AD involve object recognition and not visuospatial tasks. In fact, the appearance of Balint's syndrome in AD is accompanied by very impaired object recognition in all cases (Mendez, et al., 1990).

Lesion Method / NUMBER OF EXPERIMENTS		Attenuate	Delete	Noise	TOTAL
Weight Selection Method	All	5	7	5	17
	Unselected	2	2	1	5
	Layer	1	1	2	4
	Partition	0	0	0	2
	Value	0	2	0	2
	Sign	2	2	0	4
Unit Selection Method	All	19	6	1	26
	Unselected	4	1	0	5
	Layer	9	3	1	13
	Partition	6	2	0	8
Gain	All	2	0	0	2

Table 10: Lesion Experiments Conducted

This information, combined with the modelling results described above, has led to two hypotheses regarding lesions to the Rueckl network:

Focal Lesion Hypothesis: Focal lesions to an object recognition and spatial localization task can disproportionately affect either spatial localization or object recognition, depending on the site of damage.

Diffuse Lesion Hypothesis: Diffuse lesions to an object recognition and spatial localization task initially affect object recognition, but with sufficient damage, can disrupt spatial localization as well.

Multiple computer model lesioning experiments were conducted to test these hypotheses.

Lesioning the Rueckl Model

For this project, the (non-recurrent) feed forward connectionist network described in (Rueckl, et al., 1989) was reimplemented using the DYSNET simulator. Specific choices regarding potential functions, learning parameters, error measure, and weight updating function may differ from those in the original model (available on request). The "what" and "where" components of the network were implemented as distinct partitions of the network that share common input nodes.

Many different lesions were introduced into the model, and a simple analysis of the resulting behavior was recorded. This included (a) the number of spatial location errors (false positive and false negative); (b) the number of object identification errors; (c) the sum squared error of the "where" subnetwork; (d) the sum squared error of the "what" subnetwork; and (e) the sum squared error of the entire network.

The different lesions performed on the network are summarized in Table 10. This Table lists the number of specific experiments conducted using each selection method (e.g., selection by location or value) and specific network

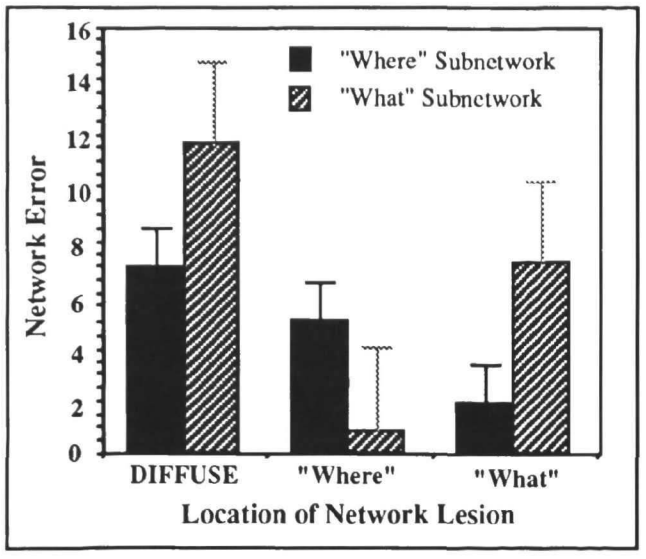

Figure 1: Lesions to Visual System Model

lesioning type (e.g., attenuation of value).

The pooled results of the experiments with focal damage ("object selection by partition" in Table 10) show that focal damage to the "parietal" (or "where") part of the network produces computational problems with the spatial orientation task and that focal damage to the "temporal" (or "what") part of the network produces problems with the object recognition part of the network. In the diffuse lesioning experiments, the pooled data show that while errors occurred in both the object recognition and spatial discrimination subtasks, the object recognition task was disproportionately affected. The results from all experiments are shown graphically in Figure 1. The sum squared error of each subnetwork across all focal and diffuse experiments was averaged and this value is shown in the graph, along with the standard error.

Discussion

The connectionist model study suggests that the neuropsychological deficits of Alzheimer's Disease and other central nervous system disruptions caused by diffuse (rather than focal) damage may have a computational basis. In other words, the particular pattern of cognitive disruption may not be due to a special predilection of the disease process for one or another part of the brain. Rather, diffuse brain injury or degeneration may cause these cognitive deficits because of the inherent nature of the brain as a distributed computational device, with patterns of connectivity that implement specific processing tasks.

Acknowledgements

Many thanks to Audrey Holland for many stimulating discussions leading to some of the work presented here. Thanks to the members of Pat Carpenter's and Audrey Holland's aphasia seminar for provocative conversations. Special thanks to Gloria Hoffman and Mark Fitzsimmons for discussions, support, and manuscript suggestions.

References

Balint, R. 1909. Seelenlähmung des 'Schauens', Optische Ataxie, raümliche Störung der Aufmerksamkeit. *Monatsschr Psychiatr Neurol* 25:51-81.

Boller, F.; and Grafman, J. ed. 1990. *Handbook of Neuropsychology*. Amsterdam: Elsevier Science Publishers.

Chomsky, N. 1965. Aspects of the Theory of Syntax. Cambridge, Massachusetts: The MIT Press.

Damasio, H.; and Damasio, A. R. 1989. Lesion Analysis in Neuropsychology. New York: Oxford University Press.

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

Feldman, J. A. 1989. Neural Representation and Neural Computation. In *Neural Connections, Mental Computation*, Nadel, L.; Cooper, L. A.; Culicover, P.; and Harnish, R. M. (ed.), Cambridge, Massachusetts: The MIT Press.

Galaburda, A. M.; LeMay, M.; Kemper, T. L.; and Geschwind, N. 1978. Right-Left Asymmetries in the Brain: Structural Differences between the Hemispheres may Underlie Cerebral Dominance. *Science* 199:852-856.

Mendez, M. F.; Mendez, M. A.; Martin, R.; and Smyth, K. A., Whitehouse, P.J. 1990. Complex Visual Disturbances in Alzheimer's Disease. *Neurology* 40:439-443.

Miikkulainen, R. 1990. A Distributed Feature Map Model of the Lexicon. In *Twelfth Annual Meeting of the Cognitive Science Society*, Boston, Massachusetts.

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

Newcombe, F.; and Ratcliff, G. 1989. Disorders of Visuospatial Analysis. In *Handbook of Neuropsychology*, Boller, F.; and Grafman, J. (ed.), Amsterdam: Elsevier Science Publishers.

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

Rumelhart, D. E.; and McClelland, J. L. 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Volume 1: Foundations. Cambridge, Massachusetts: The MIT Press.

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

Sejnowski, T. J.; and Churchland, P. S. 1988. Brain and Cognition. In *Foundations of Cognitive Science*, Posner, M. I. (ed.), Cambridge, Massachusetts: MIT Press.

Shepherd, G. ed. 1990. *The Synaptic Organization of the Brain*. New York: Oxford University Press.

Shepherd, G. M.; and Koch, C. 1990. Introduction to Synaptic Circuits. In *Synaptic Organization of the Brain*, Shepard, G. M. (ed.), New York: Oxford University Press.