

A Neural Model of Memory Impairment in Diffuse Cerebral Atrophy

Eytan Ruppín and James A. Reggia *

Department of Computer Science
A.V. Williams Bldg.
University of Maryland
College Park, MD 20742

May 6, 1994

Abstract

Background: Several previous computer-supported neural network models have been subjected to diffuse, progressive deletion of synapses/neurons. The goal in this past work has been to show that modeling cerebral neuropathological changes can provide a computational account for the pattern of memory degradation that occurs in diffuse degenerative processes such as Alzheimer's disease. Recently, however, it has been suggested that neural models cannot account for more detailed aspects of memory impairment, such as the relative sparing of remote versus recent memories observed experimentally in Alzheimer's disease [Carrie, 1993]. Methods: The latter claim is examined from a computational perspective, in the framework of a neural associative memory model. Results: We describe a neural network model that not only demonstrates graceful, progressive memory deterioration as progressive, diffuse network damage occurs, but also clearly exhibits differential sparing of remote versus recent memories. The model accounts for the experimentally observed temporal gradient of memory decline, the occurrence of false positive errors during retrieval, and testifies to the importance of modeling synaptic compensatory mechanisms. Conclusions: Our results show that neural models can account for a large variety of experimental phenomena characterizing memory degradation in Alzheimer patients. Specific testable predictions are generated, concerning the relation between the neuroanatomical degenerative findings and the clinical manifestations of Alzheimer disease.

*The authors are grateful to Drs. Ami Avni and David Horn for most helpful discussions. This research has been supported in part by a Rothschild Fellowship to Dr. Ruppín (M.D, Ph.D), and in part by Awards NS29414 and NS16332 from NINDS. Dr. Reggia (M.D, Ph.D), to whom correspondence should be addressed, is also with the Department of Neurology and the Institute of Advanced Computer Studies at the University of Maryland.

1 Introduction

In a recent paper in this journal, Carrie has presented a neural network model of memory degradation in diffuse cerebral atrophy [Carrie, 1993]. Carrie's work is an important effort in the investigation of the potential of neural models to enrich our understanding of diffuse degenerative syndromes such as Alzheimer's disease. Such models, relating the microscopic neural level of description and the macroscopic, phenomenological one, are naturally suited for the investigation of the relation between neuropathological changes and the clinical presentation of neurodegenerative diseases.

Carrie's paper provides a thorough discussion of the relevance of neural modeling to investigating memory degradation in diffuse cerebral atrophy. He finds that while a neural network model can capture the gradual, progressive degradation observed in Alzheimer's disease (e.g., [Katzman, 1986, Katzman, 1988, Drachman *et al.*, 1990], it cannot capture the relative sparing of earlier memories that has been revealed by several neuropsychological studies (e.g., [Beatty *et al.*, 1988, Kopelman, 1989]. Our work reconsiders these findings, and continues their investigation in a more biologically realistic neural model.

In his simulations, Carrie uses an early version [Hopfield, 1982] of what are currently referred to as *attractor neural networks* [Amit, 1989]. An attractor neural network is an assembly of model neurons connected recurrently by synapses (see Figure 1). The network's state is repeatedly updated; when a neuron fires, its output, weighted by synaptic strengths, is communicated to the neurons to which it is connected. This spreading activity serves as input to those neighboring neurons, and may, in turn, trigger those other neurons to fire. By using specific learning rules that govern the way synaptic strengths in the network are established, a specific set of input patterns can be memorized or stored in the network, i.e., made to be 'attractors' of the network dynamics. The term attractor here means that, if a pattern which is sufficiently similar to one of the stored memory patterns is presented as input to the network, the network's state will gradually evolve until it converges to the state representing that memory pattern. Such a network may therefore be regarded as an associative memory system.

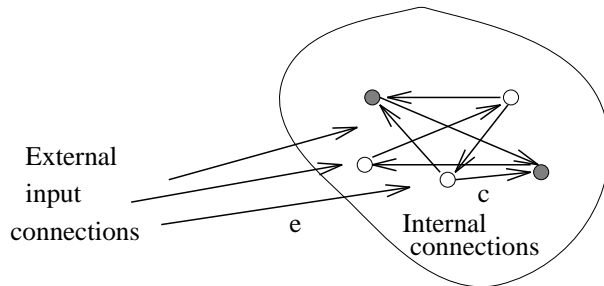


Figure 1: A schematic illustration of an attractor neural network; firing (gray) and quiescent (white) neurons, and their internal (c) and external (e) synaptic connections.

In attractor neural networks, stored memories are not represented locally at specific neurons of the network, but their corresponding representations are distributed; many neurons participate in a given memorized pattern, and a particular neuron participates in several different patterns. Representing stored memories as attractors corresponds to the intuitive notion of the persistence of cognitive concepts along some temporal span. It also is supported by biological findings of delayed, post-stimulus, sustained activity in memory-related tasks [Fuster and Jervey, 1982, Miyashita and Chang, 1988]. These experiments show that in the temporal cortex there are sustained, elevated spike rate distributions that persist for a few seconds after the removal of the stimulus. These reverberations are observed in localized modules, each about $1mm$ in diameter, holding approximately 10^5 neurons. This persistent activity is not a single neuron property, but reflects a collective behavior. [Grinasty *et al.*, 1993] have shown that many of the phenomena observed by Miyashita and Chang [1988] can be captured in an attractor neural model; they have successfully reproduced the conversion of the temporal order of the learned patterns into the spatial correlations found between them.

In his paper, Carrie reaches two primary conclusions based on a series of computer simulations with lesioned attractor neural networks:

1. Carrie finds that such models show a *gradual degradation* in their memory retrieval performance as the number of viable neurons (and their synapses) is decreased. The observed relationship was close to linear. This result is surprising because recent theoretical results indicate that attractor neural networks should continue to perform well during progressive lesioning until a critical stage is reached and rapid deterioration

occurs in memory performance [Amit *et al.*, 1985, Amit, 1989, Koscielny-Bunde, 1990]. This difference has important clinical implications, since the theoretical results imply that gradual, diffuse cortical atrophy should lead to a sudden and sharp decline of memory capacities, in large, cortical-like networks.

2. Carrie finds that ‘recent’ and ‘remote’ (old) memories, i.e., memories stored after or before network lesioning, respectively, are recalled equally well by a damaged network. Such a result is inconsistent with relevant psychological experimental studies, which demonstrate a *relative sparing of remote memories* in patients with diffuse cerebral atrophy [Beatty *et al.*, 1988, Kopelman, 1989]. Carrie attributes this failure to the simplicity and homogeneity of the attractor networks studied relative to biological systems.

In this paper, we repeat Carrie’s simulations in substantially larger networks, based on a more biologically realistic model. Under these conditions we show that while there is an increase in memory impairment with increasing lesion severity (Lashley’s mass effect [Wood, 1978]), this impairment is not the smooth, near-linear process suggested by Carrie. The clinically observed gradual deterioration of memory, however, can still be achieved if synaptic compensatory changes that accompany neural degradation are taken into account. These *synaptic compensatory changes*, manifested by an increase of the synaptic size of the remaining synapses, have been recently reported in several neuroanatomical morphometric studies of Alzheimer’s disease [Bertoni-Freddari *et al.*, 1988, Bertoni-Freddari *et al.*, 1990, DeKosky and Scheff, 1990]. When these compensatory changes are modeled by strengthening the retained synaptic connections in the network, a gradual rather than sudden decline of retrieval performance is achieved even in large neural networks, as shown in [Horn *et al.*, 1993]. By modeling memory storage in a more biologically realistic fashion as an activity-dependent, local process (versus the one-shot learning commonly used; see next section), we show that remote or pre-lesioning memories are relatively spared compared to recent memories.

In light of findings that the extent of neuronal loss in Alzheimer’s disease is less than 10% even at advanced stages [Katzman, 1986], but that the synapse to neuron ratio is significantly decreased [Davies *et al.*, 1987, Bertoni-Freddari *et al.*, 1990], we repeat Carrie’s simulations, but also do additional simulations where we retain the neurons and randomly delete some fraction of their synapses. The results demonstrate that the factor determining

the network’s performance is essentially the synapse-to-neuron ratio. Extending further on Carrie’s experiments, we examine the retrieval of previously stored memories in more detail. We show that our model can also account for experimental psychological data demonstrating a temporal gradient with relative sparing of remotely-stored memories [Beatty *et al.*, 1988, Sagar *et al.*, 1988, Kopelman, 1989]. In addition, we examine the response of the network to distractors (i.e., non-stored input patterns), and find that the networks behavior parallels the experimental finding of a significantly higher rate of false positive responses in Alzheimer’s patients versus normal controls [Kopelman, 1985]. As will be shown, both these results are achieved in the framework of our model only if synaptic compensatory changes are incorporated.

2 Methods

We use a variant of Hopfield’s attractor neural network model, proposed by [Tsodyks and Feigel’man, 1988]. Each neuron i is described by a binary variable $S_i = \{1, 0\}$ denoting an active (firing) or passive (quiescent) state, respectively. All N neurons in the network have a fixed uniform threshold θ , whose value is determined such that it yields the best memory retrieval performance in the initial undamaged state [Horn *et al.*, 1993]. Each neuron receives on the average $K \leq N$ incoming randomly determined synapses, and M memories are stored in the network. The neuron’s state is updated stochastically, in accordance with its inputs, i.e., the signals it receives from its neighbors and from external sources.

A more detailed formal description of the model is presented in the Appendix. As illustrated in Figure 1, the neurons receive two kinds of connections: 1. *external connections*, via which external input patterns are presented to the network, and 2. *internal connections*, which store the memorized patterns and whose synaptic strengths may change as a function of the neural activity in the network. The network has two behavioral modes:

1. In the *learning (storage) mode*, an input pattern, which is to be memorized, is presented to the network via the external synaptic inputs and gradually it is engraved on the internal synaptic matrix via Hebbian activity-dependent synaptic changes. Our network is hence a *repetitive-learning* system, as a pattern to be stored must be presented to the network several times before it becomes engraved on the synaptic matrix with sufficient strength, and is not simply enforced on the network in a ‘one-shot’ learning process (as, e.g., in Carrie’s simulations [1993]) (This is one way in which

our model is more ‘biologically-realistic’ than Carrie’s [1993]; see last paragraph in this section. From a computational point of view, repetitive-learning networks utilize the attractor properties of the network not only for error-correction (as in the case of one-shot learning networks) but also for learning. They are also probably more biologically plausible as they are more suited for the storage of input patterns in the presence of errors and noise [Parisi and Nicolis, 1990]). Several patterns ξ^μ may be stored in this manner as they are presented sequentially to the network, where superscript μ indicates a pattern index. Each of the N elements of a given memory pattern are chosen to be 1 (0) with probability p ($1 - p$) respectively, with $p = 0.1$ in all of the simulations presented here.

2. In the *retrieval* mode, an input pattern is presented to the network via the external connections. This pattern, which is a weakened (or distorted) version of one of the stored memory patterns, serves as a cue with which the associative memory is probed. In the network’s undamaged state, if the cued input pattern is sufficiently similar to one of the stored memories, the network will converge to a state highly similar to that memory. However, if the input pattern is not sufficiently similar, or if the network is damaged to the extent that its retrieval capacities are lost, the network will most likely remain wandering around in its initial autonomous state of random, near-zero, activity.

In its initial, undamaged state, the parameters determining the strength of the external projections and the neuronal threshold are chosen such that the network will flow dynamically into the cued memory patterns with high probability. Pathological changes are modeled either by randomly eliminating neurons from the network together with their connections, or by leaving the neurons intact but randomly deleting a fraction of their synapses.

The performance of the network is measured as follows. Suppose that a pattern ξ^μ is presented as an external input to the network (either during learning or during retrieval). After the network has converged to a stable state, or after a certain amount of time has elapsed if the network does not converge to a stable state, we measure the *overlap* m^μ achieved between the input pattern ξ^μ and the current state S of the network, defined by

$$m^\mu(t) = \frac{1}{p(1-p)N} \sum_{i=1}^N (\xi_i^\mu - p) S_i(t) . \quad (1)$$

This conventionally used *overlap* similarity measure [Tsodyks and Feigel’man, 1988, Tsodyks,

1988], much like Carrie’s similarity measure, keeps track of the neurons which should correctly fire (i.e., are turned on in the cued pattern), but it also counts with lower weighting the erroneously firing ones. The performance of the network is the average overlap obtained over all trials performed in a given architecture.

As is evident, the Tsodyks & Feigel’man model used here has several features which are more biologically-realistic than the Hopfield model used by Carrie. The transformation, from the symmetric $\{-1, +1\}$ to the asymmetric $\{0, 1\}$ notation is not at all the trivial change it may initially seem. First, unlike in the Hopfield model, there is a population of neurons that is truly quiescent, i.e., does not transfer information to the network. Second, the network processes low-activity neural firing patterns, like those reported in the cortex (e.g., [Abeles *et al.*, 1990]). Third, activity-dependent synaptic changes occur primarily between firing neurons, while the magnitude of synaptic modification occurring between two connected quiescent neurons is much smaller. But most important, as shown in [Horn *et al.*, 1993], more realistic modeling of synaptic changes requires neurons with a *non-zero* positive threshold (reflecting the difference between the resting potential and the firing threshold).

3 Experiments and Results

We have performed five types of simulations, examining the memory performance of the network as it is progressively damaged. We have studied the resulting pattern of memory decline manifested, the differential sparing of remote memories, the temporal gradient of memory retrieval and the decrease in retrieval specificity. Unless specified otherwise, the parameter values used in all experiments are (see Appendix for details) $N = 400$, $M = 20$, $K = 400$, $p = 0.1$, $\theta = 0.048$, $T = 0.005$, $\gamma = 0.025$, $e_l = 0.065$ and $e_r = 0.035$.

Experiment 1.

To track down the pattern of memory decline resulting from diffuse neural and synaptic damage, we stored 20 randomly-generated memory patterns and repeated the simulation reported in [Carrie, 1993]. in this larger fully-connected network, studying the pattern of memory decline as neurons and their synapses are gradually randomly deleted (in steps of 20 neurons each time). The stored memory patterns were presented as external inputs to the network, and its performance in each trial was determined by the final overlap achieved between the network’s stable state and the cued memory pattern, measured for

the viable, non-damaged neurons. All memory patterns were stored in the intact network, and no learning occurred during this lesion-measurement process. The results, presented in Figure 2, report the average final overlap achieved in the 50 trials performed for each level of neural deletion. A sharp performance deterioration, occurring at some narrow critical region of neuronal deletion, is evident, in contrast with the typical clinical gradual pattern.

The same process of gradual neural deletion was then repeated, but this time a *variable* synaptic compensation strategy was incorporated, where the magnitude of the remaining synapses is uniformly strengthened in a manner that partially compensates for the decrease in the neuron's input field. This procedure is motivated by recent neuroanatomical investigations in Alzheimer patients which have found an increase of the remaining synaptic size, taking place concomitantly with the degenerative (neural and synaptic) deletion processes which considerably decrease the synapse to neuron ratio [Bertoni-Freddari *et al.*, 1990, DeKosky and Scheff, 1990]. For every fraction of deletion $d \in (0, 1)$, synaptic compensation is employed by multiplying each remaining synapse's magnitude by c , with $c = 1 + \frac{dk}{1-d}$ and $k = 0.1 + 0.25d$. As evident, while synaptic compensation is initially low, it gradually increases as damage advances, reflecting an increasing effort to maintain memory capacities as the danger of functional collapse becomes eminent (the interested reader may refer to [Horn *et al.*, 1993] for a detailed technical explanation of this issue). As shown in Figure 2 (with synaptic compensation), a more gradual performance degradation to near-zero levels now occurs along a broader span of synaptic deletion than in the compensatory-absent case.

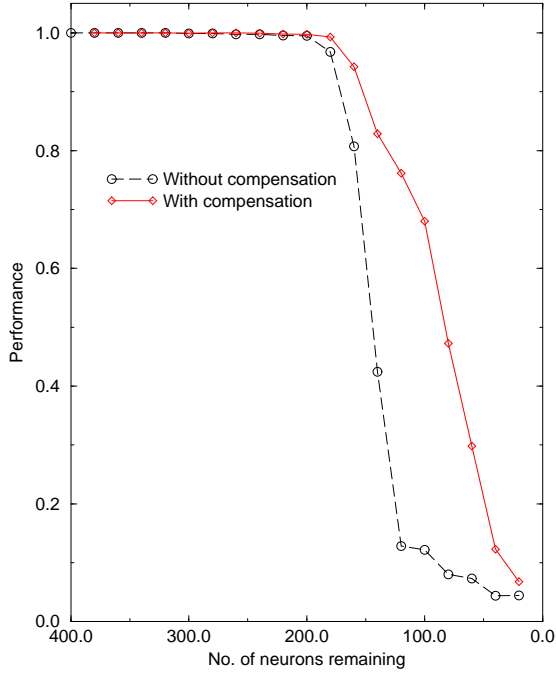


Figure 2: Post-damage retrieval performance versus neural loss.

Experiment 2.

To examine the retrieval of remote versus recently stored memories, a more complex protocol was required, involving memory retrieval and storage both before and after neural damage has occurred. This latter type of experiment is composed of five different phases, as follows:

1. **Premorbid learning** - 20 memory patterns were stored in an intact network of 400 neurons and K incoming synapses per neuron. The strength of the external input synapses was sufficient to ensure that these patterns are learned almost perfectly - i.e., that the final overlap with the learned patterns (averaged over 50 trials) was greater than 0.97. The retrieval acuity for these patterns was then verified by presenting them as cues to the network (i.e., with weaker synaptic strength) and measuring the network's performance.
2. **Diffuse damage** - After this premorbid baseline has been achieved, synapses (and possibly neurons) were deleted, leaving each remaining neuron with some $K' < K$ incoming synapses. The network performance was reexamined, as described in the next three phases.
3. **Post-damage learning** - After damage was inflicted, a single additional set of 20

different memory patterns was learned.

4. **Post-damage retrieval of remote patterns** - The recall of the previously stored memories was then reexamined.
5. **Post-damage retrieval of recent patterns** - The recall of the patterns stored after the damage occurred was examined.

Following Carrie’s experiment, we first examined recent versus remote memory in a fully-connected network whose neurons have been gradually damaged, but in the more realistic process described in the methods section, where input patterns are learned via synaptic activity-dependent changes and not simply by one-shot learning. In all of the following simulations, the network’s performance in the pre-damage state was almost perfect, and we present the retrieval of remote versus recent memories in the post-damage phase without synaptic compensation. The results of this simulation, displayed in Figure 3, demonstrate the lower retrieval of recent versus remote memories.

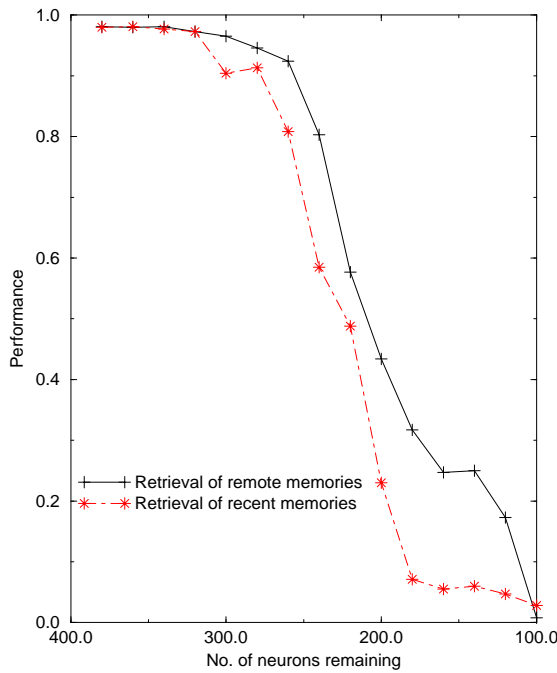


Figure 3: Post-damage retrieval performance versus neural loss - after post-damage memory storage. No synaptic compensation is used.

We repeated this experiment, but this time instead of deleting neurons we removed synapses (starting from a premorbid connectivity of $K = 200$). Figure 4 presents the results of a simulation where memory retrieval performance is measured as synaptic connections

are gradually deleted in a random manner, leaving the neurons otherwise intact. As in the previous case, the retrieval of remote memories is superior to the retrieval of recently stored ones.

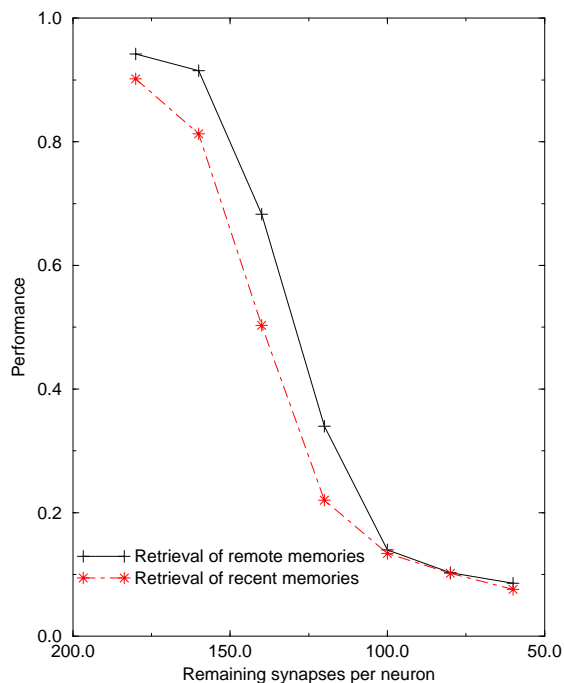


Figure 4: Post-damage retrieval performance versus synaptic loss. Remote and recent memory retrieval are examined after post-damage memory storage, without synaptic compensation, ($K = 200$).

Figure 5 presents the results of an experiment similar to that shown in Figure 4, except that now synaptic deletion is counteracted by synaptic compensation. A *fixed* synaptic compensation strategy is performed, such that the neuron’s average membrane potential (field) retains its pre-damage values and the neuronal threshold remains in its optimal value (see [Horn *et al.*, 1993]). As in the case of no synaptic compensation (during the post-damage period), recent memory retrieval is impaired earlier and more severely than remote memory retrieval. However, incorporating synaptic compensatory changes results in an earlier onset of decreased recent memory than seen without compensation (Figure 3), and further increases the gap between recent and remote memory retrieval. Repeating this simulation, but with variable instead of fixed compensation, yields a similar decrease of recent memory retrieval.

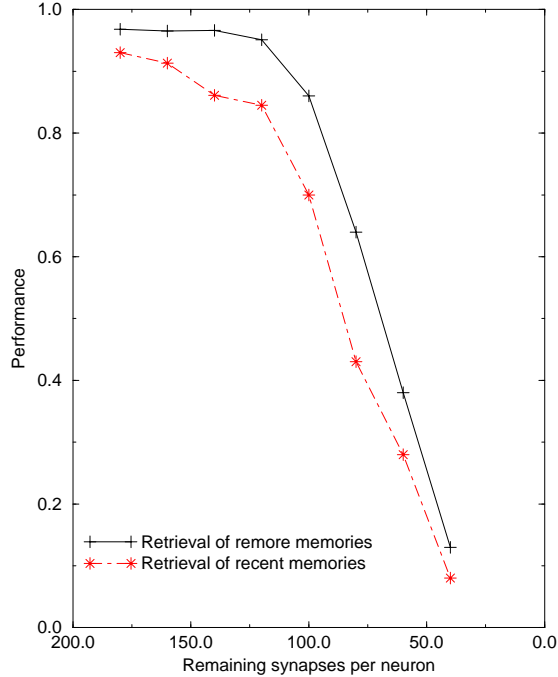


Figure 5: Post-damage retrieval performance with synaptic compensation - after post-damage memory storage.

Experiment 3.

To examine the temporal gradient of memory retrieval reported in the psychological literature, twenty-five memory patterns were arbitrarily divided into five sets, each having five memories. Then, five subsequent steps of synaptic degeneration and memory storage were performed; in each such step 25 of the currently existing incoming synapses were deleted from each neuron, and a set of five memories was then presented and stored in the network in an activity-dependent manner. So, for example, while the first set of memories was stored with a rather intact network (average connectivity of 175 per neuron) the last set was stored in a damaged network that had lost more than half of its original internal connectivity (average connectivity of 75 per neuron). This storage process simulates memory storage over a period of time, in the presence of continued diffuse loss of synaptic connections. After this process was completed, we studied the retrieval of memories from each of the stored sets using the current state of the network (i.e., with average connectivity of 75 synapses per neuron), and examined the temporal gradient achieved. The results of this experiment are presented in Figure 6. The memory set numbers are displayed on the x-axis, such that set No. 1 denotes the most remotely stored memories and set No. 5 the most recently stored set. As shown, with no synaptic compensation the (low) performance has no temporal

gradient. (This absence of a temporal gradient remains true also in simulations run with lesser levels of damage where the performance achieved is higher, but these results are not shown here. With very little synaptic degeneration and no compensation, a slight inverse temporal gradient is observed, where the performance of most recent memories is highest.) With synaptic compensation, the experimentally observed temporal gradient is obtained for a wide range of synaptic damage. These results are demonstrated both for the case where the external input is the cued pattern applied via a weak external field (A), and for the case where the external input is a subset of the cued pattern (B) (see Appendix). As illustrated, these two ways of input presentation yield similar results.

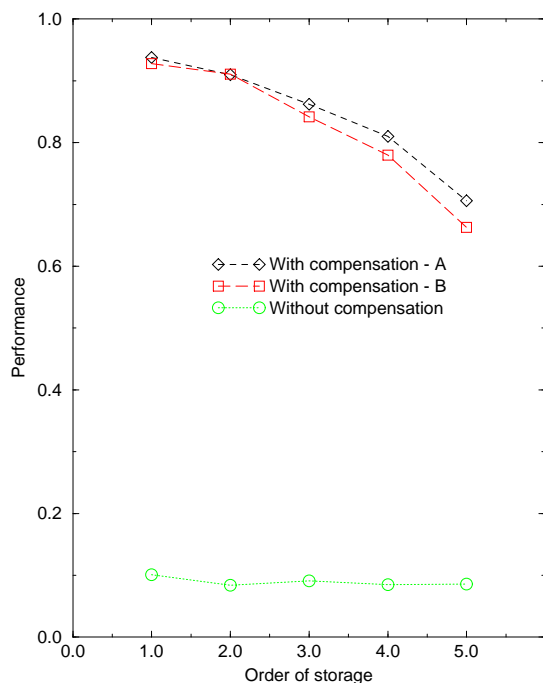


Figure 6: Retrieval performance of memories stored at different levels of synaptic degeneration. In case B, where the external input was applied as a subset of the correct memory (only 80% of the input neurons that should be on in the input pattern actually fire), $e_r = e_l = 0.065$.

Experiment 4.

To examine the rate of false positive response, non-input patterns were presented to the network. The novel (not previously stored) external input patterns were randomly generated, with an average activity level of 0.05. When presented with such an input pattern, the network may either remain in its low activity basal state (interpreted as a correct, non-response reaction), or it may converge to a stable state possibly having high overlap with

one or more memory patterns (interpreted as a false positive, erroneous ‘recognition’ of a non-stored pattern). Starting from an intact network, we repeatedly measured the highest overlap achieved with any of the stored patterns (averaged over many trials), as the level of diffuse damage in the network is increased. Figure 7 shows that synaptic compensation significantly increases the likelihood of false positive errors in a network undergoing synaptic degeneration. It is of interest to note that the network may have a considerable false positive error rate, while still maintaining a high level of memory retrieval (compare with Figure 5). Note also that as the level of diffuse synaptic damage increases beyond some point the false positive error rate begins to taper off. This is characteristic of networks employing synaptic compensation, as will be discussed below.

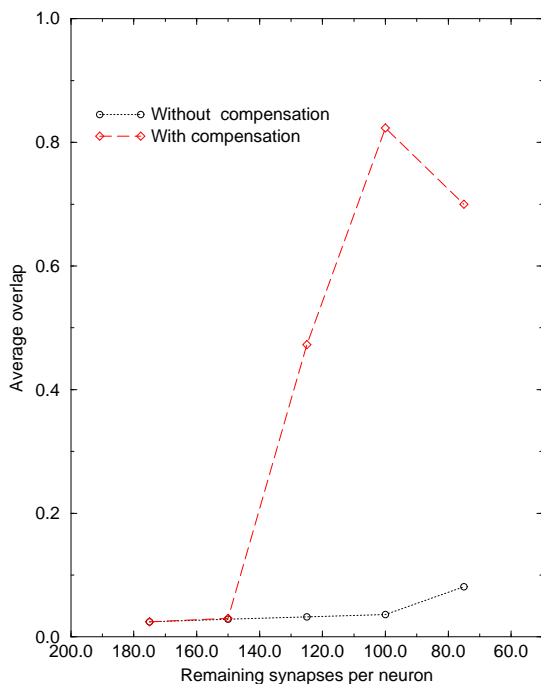


Figure 7: False positive error rate at different levels of synaptic degeneration.

Experiment 5

The results of experiment 2 (Figure 3) suggest that failed memory storage (yielding poor recent memory) may accelerate the onset of the degradation of remote memory performance. This raises the possibility that even re-learning of memory patterns that have been already stored in the network may cause memory degradation at advanced levels of damage. To examine this issue, we investigated memory retrieval performance at progressing levels of synaptic deletion (with synaptic compensation), comparing the performance without re-

learning with that achieved with re-learning. Re-learning was performed by presenting the remote memory patterns (i.e., the *same* memory patterns stored in the intact, pre-lesion network) again to the damaged network. The results, shown in Figure 8, demonstrate that while at low levels of damage re-learning improves performance, at high levels of damage it actually worsens it.

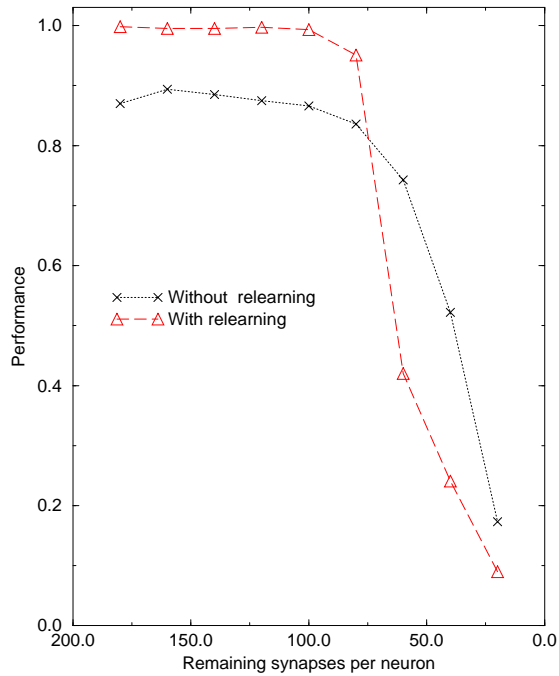


Figure 8: The effect of re-learning on memory performance. $\gamma = 0.02$.

4 Discussion

During the last several years there has been a considerable increase of interest in using neural models to explore potential pathophysiological theories of psychiatric and neuropsychological disorders (see [Reggia *et al.*, 1994 in press], for a recent review). In this paper we have specifically examined how attractor neural network models can qualitatively account for four basic features of memory degradation in diffuse cerebral atrophy: the gradual pattern of ‘graceful’ memory deterioration; the differential sparing of old memories versus recent ones; the typical temporal gradient of memory decline, and the increased rate of false positive errors in retrieval. From the simulations described above, several conclusions can be made:

- In an attractor neural network of considerable size (above a few hundreds of neurons) undergoing progressive, diffuse loss of neurons/synapses without synaptic compensation, memory retrieval is initially preserved until a critical point is reached where retrieval sharply declines and remains very poor thereafter. We believe that the gradual, almost linear, pattern of decline obtained in an earlier study [Carrie, 1993] was due to the very small (100 neurons) and sparsely loaded (1 or 2 memories stored) network used in that study.
- Nevertheless, the clinically gradual pattern of memory degradation in Alzheimer’s patients can be simulated if synaptic compensatory changes are incorporated into the model. More specifically, gradual decline is obtained with a variable synaptic compensation strategy, where the compensatory effort increases in a faster rate than the progression of the neurodegenerative damage.
- Both of these conclusions remain true either when synaptic deletion is performed in a random manner, with no neural deletion (as shown in [Horn *et al.*, 1993]), or when synaptic deletion is accompanying neural damage.
- Remote memory is relatively spared when compared to recent memory, paralleling recent results obtained with Alzheimer’s patients. Synaptic compensation further increases the degradation of recently-stored memories. As we shall discuss further on, these findings are true in general (i.e., across a very broad parameter range) when synaptic compensation is incorporated. But without synaptic compensation the situation is complex, and the network’s behavior is sensitive to the choice of parameters.

We believe that the inability to demonstrate this phenomenon in the one previous neural modelling study that looked for it [Carrie, 1993] arose from ignoring synaptic compensatory changes, and from the non-biological nature of the one-shot synaptic storage method that was used.

- Synaptic compensation must be incorporated into the model in order to simulate the temporal gradient of memory retrieval and the increased false positive error rates reported experimentally in Alzheimer’s patients.
- At high levels of damage, an attempt to learn additional patterns may result in the degradation of remote memory. This is true even when presenting the already stored memories as input patterns to the network.

Simple signal-to-noise considerations show that the relative sparing of remote memories is a general characteristic of our model and not just a phenomenon that manifests itself in some particular narrow range of parameters. This is true only if patterns are stored via a repetitive-learning rule like that employed in this work (which makes memory storage more liable to damage than memory retrieval), and only as long as synaptic compensation takes place along with synaptic deletion. The situation without synaptic compensation is much more complex and the relative sparing of remote memories may not be observed. Hence, the model provides a solid account of the temporal gradient observed clinically, extending back to the time of onset of the degenerative processes. However, the clinical temporal gradient may extend back to childhood and early adulthood, that is, long before the degenerative processes are assumed to take place [Kopelman, 1989]. In light of this data, it has been previously suggested that early memories are preserved because they are well rehearsed [Kopelman, 1989]. We examined this hypothesis in a simulation similar to that described in Experiment 3, except that some additional memories were stored prior to the onset of damage. In each storage epoch, in addition to the learning of new patterns, a few of the patterns already stored were presented again as inputs to the network, enforcing their storage. As before, a marked temporal gradient was achieved, but now extending back to the period before the onset of damage, as memories that were rehearsed more are better preserved.

Several interesting related issues remain open for further investigation. One task is to explain why the temporal gradient in Alzheimer is less steep than in Korsakoff patients [Kopelman, 1989]. To this end, however, one needs to develop a spatially organized model,

where the distinction between diffuse and focal lesions is meaningful. Second, it would be useful to examine computationally the recently published anatomical models of memory impairment in Alzheimer's disease, either due to frontal dysfunction [Kopelman, 1991], or due to bilateral damage to the anterior temporal lobes [Kapur *et al.*, 1992]. For example, [Ruppin *et al.*, 1994] have recently studied the possibility of modeling Stevens' theory that the onset of schizophrenia is associated with regenerative synaptic changes occurring in the frontal lobes after the degeneration of incoming temporal lobe projections. Thirdly, while this work has concentrated on the study of episodic memory, the modeling of the semantic aspects of memory in Alzheimer's disease is obviously warranted. A preliminary attempt in this direction has already been presented by [Herrmann *et al.*, 1993].

Finally, we conclude by discussing some implications and predictions:

- The model predicts that Alzheimer's patients with the same level of neurodegenerative changes, but differing in the level of synaptic compensation, will exhibit two distinct manifestations of memory disturbances. Those patients with minor compensatory synaptic changes should suffer from a considerable, general decrease in memory retrieval, but should have relatively maintained learning capacities and hence, preserved recent memory (an 'inverse' temporal gradient). Those patients with considerable synaptic compensatory changes should have relatively maintained remote memory, but decreased learning capacities leading to a deteriorated recent memory (and the temporal gradient observed typically). These predictions are in fact testable, in light of the recent developments in morphometric techniques which have enabled detailed studies of the neurodegenerative changes accompanying Alzheimer's disease, not only in autopsies (e.g. [Bertoni-Freddari *et al.*, 1988]) but also in cortical biopsies [DeKosky and Scheff, 1990].
- The finding that post-damage learning may actually impair remote memory retrieval may have surprising clinical implications: As long as neurodegenerative damage is small, re-learning of stored patterns will improve their recognition, as is naturally expected. However, beyond some level of neurodegenerative damage, when learning becomes impaired, retraining previously stored patterns may paradoxically be harmful. These results are in accordance with the clinical findings that gains are small or nonexistent when traditional memory training procedures are used in intervention research with Alzheimer patients, unless conceptual support is provided both for en-

coding and retrieval [Herlitz *et al.*, 1992] (such support presumably delays the learning impairment).

We have seen that memory degradation in Alzheimer's disease is not just a simple, 'monotone' function of neurodegenerative changes, but is a result of the combination of these changes and counteracting synaptic compensatory changes. This may contribute to the inconsistency in studies searching for a correlation between overall memory performance and a gross parameter of brain damage such as cortical atrophy (e.g., [Dall'Ora *et al.*, 1989] versus [Kopelman, 1989]). It is our hope that neural models may have an important role in finding a relationship between more detailed, microscopic neuropathological changes and the clinical manifestations of brain diseases.

References

- [Abeles *et al.*, 1990] M. Abeles, E. Vaadia, and H. Bergman. Firing patterns of single units in the prefrontal cortex and neural network models. *Network*, 1:13, 1990.
- [Amit *et al.*, 1985] D. J. Amit, H. Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55:1530–1533, 1985.
- [Amit, 1989] D.J. Amit. *Modeling brain function: the world of attractor neural networks*. Cambridge University Press, 1989.
- [Beatty *et al.*, 1988] W. Beatty, D.P. Salmon, N. Butters, W.C. Heindel, and E.L. Granholm. Retrograde amnesia in patient's with Alzheimer's disease or Huntington's disease. *Neurobiology of Aging*, 9:181–186, 1988.
- [Bertoni-Freddari *et al.*, 1988] C. Bertoni-Freddari, W. Meier-Ruge, and J. Ulrich. Quantitative morphology of synaptic plasticity in the aging brain. *Scanning Microsc.*, 2:1027–1034, 1988.
- [Bertoni-Freddari *et al.*, 1990] C. Bertoni-Freddari, P. Fattoretti, T. Casoli, W. Meier-Ruge, and J. Ulrich. Morphological adaptive response of the synaptic junctional zones in the human dentate gyrus during aging and Alzheimer's disease. *Brain Research*, 517:69–75, 1990.
- [Carrie, 1993] J.R. Carrie. Evaluation of a neural network model of amnesia in diffuse cerebral atrophy. *British Journal of Psychiatry*, 163:217–222, 1993.
- [Dall'Ora *et al.*, 1989] P. Dall'Ora, S. Della Sala, and H. Spinnler. Autobiographical memory: Its impairment in amnesic syndromes. *Cortex*, 25:197–217, 1989.
- [Davies *et al.*, 1987] C.A. Davies, D.M.A. Mann, P.Q. Sumpter, and P.O. Yates. A quantitative morphometric analysis of the neuronal and synaptic content of frontal and temporal cortex in patient with Alzheimer's disease. *J. Neurol. Sci.*, 78:151–164, 1987.
- [DeKosky and Scheff, 1990] S. T. DeKosky and S.W. Scheff. Synapse loss in frontal cortex biopsies in Alzheimer's disease: Correlation with cognitive severity. *Ann. Neurology*, 27(5):457–464, 1990.

- [Drachman *et al.*, 1990] D. A. Drachman, B. F. O'Donnell, R. A. Lew, and J.M. Swearer. The prognosis in Alzheimer's disease. *Arch. Neurol.*, 47:851–856, 1990.
- [Fuster and Jervey, 1982] J.M. Fuster and J.P. Jervey. Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. *The Journal of Neuroscience*, 2(3):361–375, 1982.
- [Griniasty *et al.*, 1993] M. Griniasty, M.V. Tsodyks, and D.J. Amit. Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Computation*, 5:1–17, 1993.
- [Hasselmo, 1993] M.E. Hasselmo. Acetylcholine and learning in a cortical associative memory. *Neural computation*, 5(1):32, 1993.
- [Herlitz *et al.*, 1992] A. Herlitz, B. Lipinska, and L. Backman. Utilization of cognitive support for episodic remembering in alzheimer's disease. In L. Backman, editor, *Memory functioning in Dementia*, pages 73–95. North-Holland, 1992.
- [Herrmann *et al.*, 1993] M. Herrmann, E. Ruppin, and M. Usher. A neural model of the dynamic activation of memory. *Biological Cybernetics*, 68:455–463, 1993.
- [Hopfield, 1982] J.J. Hopfield. Neural networks and physical systems with emergent collective abilities. *Proc. Nat. Acad. Sci. USA*, 79:2554, 1982.
- [Horn *et al.*, 1993] D. Horn, E. Ruppin, M. Usher, and M. Herrmann. Neural network modeling of memory deterioration in Alzheimer's disease. *Neural Computation*, 5:736–749, 1993.
- [Kapur *et al.*, 1992] N. Kapur, D. Ellison, M.P. Smith, D.L. McLellan, and E.H. Burrows. Focal retrograde amnesia following bilateral temporal lobe pathology. *Brain*, 115:73–85, 1992.
- [Katzman, 1986] R. Katzman. Alzheimer's disease. *New England Journal of Medicine*, 314(15):964–973, 1986.
- [Katzman, 1988] R. Katzman. Comparison of rate of annual change of mental status score in four independent studies of patients with Alzheimer's disease. *Ann. Neurology*, 24(3):384–389, 1988.

- [Kopelman, 1985] M.D. Kopelman. Rates of forgetting in Alzheimer-type dementia and Korsakoff's syndrome. *Neuropsychologia*, 23:623–638, 1985.
- [Kopelman, 1989] M.D. Kopelman. Remote and autobiographical memory, temporal context memory, and frontal atrophy in Korsakof and Alzheimer patients. *Neuropsychologia*, 27:437–460, 1989.
- [Kopelman, 1991] M.D. Kopelman. Frontal dysfunction and memory deficits in the alcoholic Korsakoff syndrome and Alzheimer-type dementia. *Brain*, 114:117–137, 1991.
- [Koscielny-Bunde, 1990] E. Koscielny-Bunde. Effect of damage in neural networks. *Journal of Statistical Physics*, 58:1257 – 1266, 1990.
- [Miyashita and Chang, 1988] Y. Miyashita and H.S. Chang. Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature*, 331:68–71, 1988.
- [Parisi and Nicolis, 1990] D.J. Amit D.J. and G. Parisi and S. Nicolis. Neural potentials as stimuli for attractor neural networks. *Network*, 1:75–88, 1990.
- [Reggia *et al.*, 1994 in press] J. Reggia, R. Berndt, and L. D'Autrechy. Connectionist models in neuropsychology. In *Handbook of Neuropsychology*, volume 9. 1994, in press.
- [Ruppin *et al.*, 1994] E. Ruppin, J. Reggia, and D. Horn. Compensatory mechanisms in an attractor neural network model of schizophrenia. *Neural Computation*, page In Press, 1994.
- [Sagar *et al.*, 1988] H.J. Sagar, N.J. Cohen, E.V. Sullivan, S. Corkin, and J.H. Growdon. Remote memory function in Alzheimer's disease and Parkinson's disease. *Brain*, 111:185–206, 1988.
- [Tsodyks and Feigel'man, 1988] M.V. Tsodyks and M.V. Feigel'man. The enhanced storage capacity in neural networks with low activity level. *Europhys. Lett.*, 6:101 – 105, 1988.
- [Tsodyks, 1988] M.V. Tsodyks. Associative memory in assymmetric diluted network with low activity level. *Europhys. Lett.*, 7:203–208, 1988.
- [Wood, 1978] C.C. Wood. Variations on a theme by Lashley: lesion experiments on the neural model of Anderson, Silverstien, Ritz and Jones. *Psychological Review*, 85(6):582–591, 1978.

Appendix: A formal description of the model

Each neuron i is described by a binary variable $S_i = \{1, 0\}$ denoting an active (firing) or passive (quiescent) state, respectively. All N neurons in the network have a fixed uniform threshold θ . The initial state of the network $S(0)$ is random, with average activity level $q < p$, reflecting the notion that the spontaneous level of activity of a cortical network is lower than the activity level of the persistent attractor states [Miyashita and Chang, 1988]. The input (post-synaptic potential) h_i of neuron i is the sum of internal contributions from other neurons in the network and external contributions F_i^e , as sketched in Figure 1, and given by

$$h_i(t) = \sum_j W_{ij} S_j(t-1) + F_i^e \quad (2)$$

where W_{ij} is the weight of the connection from neuron j to neuron i . The updating rule for neuron i at time t is given by

$$S_i(t) = \begin{cases} 1, & \text{with probability } G(h_i(t) - \theta) \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where G is the sigmoid function $G(x) = 1/(1 + \exp(-x/T))$, and T denotes the noise level.

During *learning*, M distributed memory patterns ξ^μ are stored in the network by sequentially presenting them as inputs to the network one after the other. The elements of each memory pattern are chosen to be 1 (0) with probability p ($1-p$), respectively, with $p \ll 1$. An input pattern (say ξ^1) is memorized by orienting the external inputs F^e with it, such that

$$F_i^e = e_l \cdot \xi_i^1, \quad (e_l > 0), \quad (4)$$

where e_l is a scalar denoting the strength of the external inputs during learning. Following the dynamics defined in (2) and (3), the network state evolves until it converges to a stable state. Concomittantly, as the network iterates, the synaptic weights are modified in an activity-dependent, Hebbian-like manner, according to the rule

$$W_{ij}(t) = W_{ij}(t-1) + \frac{\gamma}{N} (\bar{S}_i - p)(\bar{S}_j - p), \quad (5)$$

where \bar{S}_k is 1 (0) only if neuron k has been firing (quiescent) for the last consecutive 5 iterations and γ is a constant determining the magnitude of activity-dependent changes. If either of the neurons i or j has not been in the same firing state for all the last 5 iterations, then the synaptic weight W_{ij} is not modified. In their initial state, the synaptic

matrix weights are taken to be zero. Every input pattern to be stored is presented 5 times to the network, and at each presentation the network is run for 10 iterations. As it is gradually engraved in the synaptic matrix, subsequent presentation of the same pattern results in stable states with increasing overlap with the learned input pattern. Eventually, the pattern is stored in the network’s synaptic matrix, and high final overlap is achieved.

Memory retrieval is modeled in a similar fashion to the way a pattern is engraved into the network, i.e., by making the input pattern be a scaled version of one of the memorized patterns (the *cued* pattern, say ξ^1), such that

$$F_i^e = e_r \cdot \xi_i^1, \quad (e_r > 0) \quad (6)$$

but e_r is typically smaller than e_l . Applying the external cue pattern via a weakened external force (i.e., assuming some form of neuromodulation that changes the synaptic strengths during learning and retrieval, as reported in [Hasselmo, 1993]) is computationally equivalent to presenting the external cue as a distorted version of one of the memory patterns, as illustrated in Figure 6. After the network reaches a stable state, its performance in a given trial is measured by the final overlap achieved between the network state and the cued memory pattern. During the recall phase, the synaptic weights are not modified. The parameters e_l and e_r are fixed and the same for all simulations in this paper.