Submitted to *NIPS 2000*.

Hippocampally-Dependent Consolidation in a Hierarchical Model of Neocortex

Szabolcs Káli **Peter Dayan**
Gatsby Computational Neuroscience Unit
17 Queen Square, London, England, WC1N 3AR.
szabolcs@gatsby.ucl.ac.uk

Abstract

In memory consolidation, declarative memories which initially require the hippocampus for their recall, ultimately become independent of it. Consolidation has been the focus of numerous experimental and qualitative modeling studies, but only little quantitative exploration. We present a consolidation model in which hierarchical connections in the cortex, that initially instantiate purely semantic information acquired through probabilistic unsupervised learning, come to instantiate episodic information as well. The hippocampus is responsible for helping complete partial input patterns before consolidation is complete, while also training the cortex to perform appropriate completion by itself.

1 Introduction

The hippocampal formation and adjacent cortical areas have long been believed to be involved in the acquisition and retrieval of long-term memory for events and other declarative information. Clinical studies in humans and animal experiments indicate that damage to these regions results in amnesia, whereby the ability to acquire new declarative memories is impaired and some of the memories acquired before the damage are lost.[1] The observation that recent memories are more likely to be lost than old memories in these cases has generally been interpreted as evidence that the role of these medial temporal lobe structures in the storage and/or retrieval of declarative memories is only temporary. In particular, several investigators have advocated the general idea that, in the course of a relatively long time period (up to decades in humans), memories are reorganized (or *consolidated*) so that memories whose successful recall initially depends on the hippocampus gradually become independent of this structure (see Refs. 2-4). However, other possible interpretations of the data have also been proposed.[5]

There have been several analyses of the computational issues underlying consolidation. There is a general consensus that memory recall involves the reinstatement of cortical activation patterns which characterize the original episodes, based only on partial or noisy input. Thus the computational goal for the memory systems is cortical pattern completion; this should be possible after just a single presentation of the particular pattern when the hippocampus is intact, and should be possible independent of the presence or absence of the hippocampus once consolidation is complete. The hippocampus plays a double role: a) supporting one-shot learning and subsequent completion of patterns in the cortical areas it is directly connected

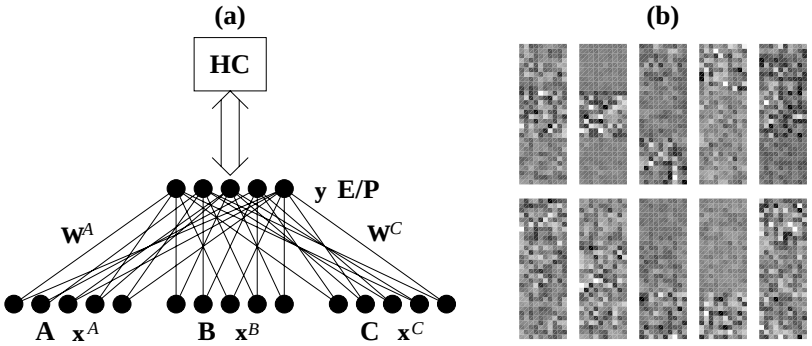

Figure 1: (a:) Model architecture. All units in neocortical areas A, B, and C are connected to all units in area E/P through bidirectional, symmetric weights, but there are no connections between units on the same level. The hippocampus (HC) is not directly implemented, but it can influence and store the patterns in E/P. (b:) Examples of the weights received by units in E/P from areas A-C after pretraining. Each rectangle represents one unit in E/P, and each little square inside the rectangles represents a weight; the color of the square indicates the magnitude of the weight (black – large positive, white – large negative). The spatial arrangement of the weights is only for visualization purposes; the weights from areas A, B, and C are shown in the top, middle, and bottom third of the rectangle, respectively.

to, and b) directing consolidation by reinstating these stored patterns in those same cortical regions and allowing the efficacies of cortical synapses to change.

Despite the popularity of the ideas outlined above, there have been surprisingly few attempts to construct quantitative models of memory consolidation. Alvarez and Squire (1994) is the only model we could find that has actually been implemented and tested quantitatively. Although it embodies the general principles above, the authors themselves acknowledge that the model has some rather serious limitations, largely due to its spartan simplicity (*eg* it only considers 2 perfectly orthogonal patterns over 2 cortical areas of 8 units each) which also makes it hard to test comprehensively. Perhaps most importantly, though (and this feature is shared with qualitative models such as Murre (1997)), the model requires some way of establishing and/or strengthening functional connections between neurons in disparate areas of neocortex (representing different aspects of the same episode) which would not normally be expected to enjoy substantial reciprocal anatomical connections.

In this paper, we consider consolidation using a model whose complexity brings to the fore consideration of computational issues that are invisible to simpler proposals. In particular, it treats cortex as a hierarchical structure, with hierarchical codes for input patterns acquired through a process of unsupervised learning. This allows us to study the relationship between coding for generic patterns, which forms a sort of semantic memory, and the coding for the specific patterns through consolidation. It also allows us to consider consolidation as happening in hierarchical connections (in which the cortex abounds) as an alternative to consolidation only between disparate areas at the same level of the hierarchy. The next section of the paper describes the model in detail and section 3 shows its performance.

2 The Model

Figure 1a shows the architecture of the model, which involves three cortical areas (A, B, and C) that represent different aspects of the world. We can understand

consolidation as follows: across the whole spectrum of possible inputs, there is structure in the activity within each area; but there are no strong correlations between the activities in different areas (these are the generic patterns referred to above). Thus, for instance, nothing in particular can be concluded about the pattern of activity in area C given just the activities in areas A and B. However, for the specific patterns that form particular episodes, there are correlations in these activities. As a result of this, it becomes possible to be much more definite about the pattern in C given activities in A and B that reinstate part of the episode. Before consolidation, information about these correlations is stored in the hippocampus and related structures; after consolidation, the information is stored directly in the weights that construct cortical representations.

The model does not assume that there are any direct connections between the cortical areas. Instead, as a closer match to the available anatomical data, we assume a hierarchy of cortical regions (in the present model having just two layers) below the hippocampus. It is hard to establish an exact correspondence between model components and anatomical regions, so we tentatively call the model region on the top of the cortical hierarchy entorhinal/parahippocampal/perirhinal area (E/P), and lump together all parts of the hippocampal formation into an entity we call hippocampus (HC). E/P is connected bidirectionally to all the cortical areas.

Semantic learning

In a hierarchical model, we have to specify the relationship between activity in E/P and that in areas A, B and C in generic or semantic circumstances (*ie* before thinking about particular episodes). Here, the model borrows from the extensive theories about the formation of hierarchical representations in cortex that comes from work in unsupervised learning (*eg* Hinton & Sejnowski, 1999). The idea is that top-down connections from E/P to A, B and C constitute a probabilistic *generative model* that captures the joint probability distribution over the activities A, B and C, *ie* the correlations between units both within and across areas. One general role for such generative models (although it is rarely described in quite this way) is to provide a statistically normative version of auto-associative memory, performing probabilistic pattern completion in a way we describe below. Indeed, we use for this hierarchical model a Restricted Boltzmann Machine (RBM),[7] which is closely related to the Hopfield net,[8] the paradigmatic auto-associative memory. However, almost any well-founded unsupervised learning model could be used instead. It is the relationship between generative models and auto-associative memory that will allow us to model semantic and episodic knowledge together.

In the RBM, activities $(\mathbf{x}^A, \mathbf{x}^B, \mathbf{x}^C, \mathbf{y})$ are binary, and weights are symmetric. Since there are no connections within the areas (or within the E/P), the dynamics of activity in the network consists of alternating updates in the two layers. At each step, units within the layers are set according to the Markov chain Monte-Carlo Gibbs sampling rule

$$x_i^A = \begin{cases} 1 & \text{with probability } \sigma\left([\mathbf{W}^A.\mathbf{y}]_i\right) \\ 0 & \text{with probability } 1-\sigma\left([\mathbf{W}^A.\mathbf{y}]_i\right) \end{cases}$$

(and similarly for the others) where $\mathbf{W}^A$ are the weights (including a bias term), and $\sigma(x) = 1/(1+\exp(-x))$ is the standard sigmoid function. All the units within a layer can be updated in parallel. Input to a layer from the world *clamps* the activities so that they do not change. Given clamped activities in $\mathbf{x}^A$ and $\mathbf{x}^B$, say, the network produce *samples* of $\mathbf{x}^C$ according to a distribution

$$P[\mathbf{x}^C|\mathbf{x}^A, \mathbf{x}^B; \mathbf{W}] \tag{1}$$

where $\mathbf{W} = \{\mathbf{W}^A, \mathbf{W}^B, \mathbf{W}^C\}$. Given appropriate weights $\mathbf{W}$, this is how a generative model performs auto-association, completing $\mathbf{x}^A, \mathbf{x}^B$ to the best fitting $\mathbf{x}^C$.

The weights $\mathbf{W}$ are assumed to be subject to slow plastic changes in order to fit distributions such as that in equation 1 to the statistics of the patterns presented. The learning rule is based on the standard Boltzmann Machine learning algorithm[9] using Gibbs sampling, with the modification that, in the negative phase, only one full step of Gibbs sampling is done.[10] This learning rule involves one phase of Hebbian learning driven by activity patterns from the world, and one phase of anti-Hebbian learning driven by patterns generated in response by the network, and is well suited for extracting the statistical regularities in the input patterns.

Episodic learning

The hippocampus is not modeled explicitly in the current model. Instead, it is assumed to be capable of three operations: (1) under appropriate conditions (*eg* when the current stimulus configuration is salient or important for some other reason), it stores a representation of the current E/P pattern, which allows the reinstatement of that E/P pattern if the corresponding hippocampal memory state is activated; (2) if the current E/P pattern is sufficiently similar to a previously stored pattern, the hippocampal representation for the stored pattern is activated (hippocampal pattern completion), and, consequently, the stored E/P pattern gets reinstated; (3) during consolidation, the patterns stored in the HC are activated intrinsically and randomly (this may happen during slow wave sleep[11]), which leads to the reinstatement of cortical patterns in the same way as during completion, at least given appropriate weights.

Prior to the end of consolidation, the hippocampus guides the process of probabilistic pattern completion by providing something like a very strong prior over the activities $\mathbf{y}$ associated with the patterns that are stored. The ultimate effect of consolidation in the model is to change the weights $\mathbf{W}$ such that they themselves embody this self-same strong prior, without the need for hippocampal input. Note that the hippocampus does not need to be able to reinstate $\mathbf{x}^A, \mathbf{x}^B$ and $\mathbf{x}^C$ directly, rather it does it using the generic knowledge encoded in the weights $\mathbf{W}$ by reinstating only the $\mathbf{y}$ pattern.

3 Simulations

In the simulation, each of the four cortical areas (A, B, C, and E/P) contained 100 units. For each of A, B, and C, we generated 10 random binary patterns (denoted $\mathbf{x}^{A_1}$–$\mathbf{x}^{A_{10}}$, $\mathbf{x}^{B_1}$–$\mathbf{x}^{B_{10}}$ and $\mathbf{x}^{C_1}$–$\mathbf{x}^{C_{10}}$, each bit of which is turned on with probability 1/2). These stand for the different stimuli that can be represented in any one of these areas. During the semantic learning phase, these patterns are presented to the network in random combinations (*eg* an input example could be $\mathbf{x}^{A_1}\mathbf{x}^{B_6}\mathbf{x}^{C_3}$), which corresponds to prior exposure of the system to events involving different combinations of stimuli. In all, 50,000 presentations were made and cortical weights were modified using the RBM learning algorithm. This leads to a population code in E/P for the patterns presented in the input layer, and establishes a correspondence between the representations in areas A-C and those in E/P in the form of a generative model. Examples of the resulting weights for E/P units are shown in Figure 1b. Some of the units seem to specialize in representing just a single input area and ignore the others, while other weights are completely global.

Next, 8 specific input patterns ($\mathbf{x}^{A_1}\mathbf{x}^{B_1}\mathbf{x}^{C_1}$–$\mathbf{x}^{A_8}\mathbf{x}^{B_8}\mathbf{x}^{C_8}$) were designated as episodic patterns to be memorized. The E/P population codes in which they result were determined and stored in the hippocampus. Finally, a consolidation phase was run. This phase consists of alternating blocks of two types of presentations. The first is identical to those in the pre-training phase, and corresponds to continued exposure to the same kinds of stimuli (while awake, or perhaps during a sleep stage

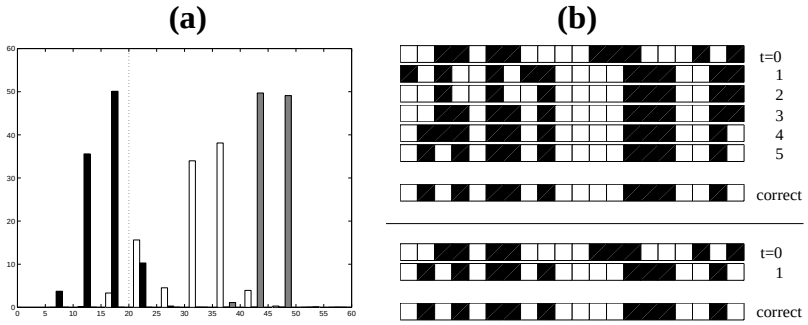

Figure 2: (a:) Threshold selection for hippocampal completion. The black bars represent the frequency histogram of the squared distances between the E/P representations of partial patterns and the correct full pattern; the gray bars are the same for the partial patterns and all the stored patterns other than the correct one; the white bars represent the distances between the E/P representations of the stored patterns and all other valid full patterns. Setting the threshold for completion of the E/P pattern to a stored pattern to 20 (dotted line) makes it possible to correctly complete 90% of partial patterns while leaving the representations of full patterns mostly unchanged. (b:) Convergence to the memorized pattern in area C (after consolidation) while a partial pattern is presented in areas A and B. The top part of the figure shows the first five iterations in the absence of the hippocampus and the correct pattern at the bottom; the bottom part shows the first iteration in the presence of the hippocampus. Only the first 20 units in area C are shown; the top row in both cases is the initial random pattern.

characterized by independent random activations of cortical areas). The second type of presentation starts from hippocampal activation of one of the stored patterns in E/P, and activation spreads to the rest of the cortex from there. Otherwise the learning procedure is unchanged. This phase also consisted of 50,000 presentations altogether, divided into 50 blocks. Within each block, 900 presentations of the second type were followed by 100 presentations of the first type.

After each phase of training, the performance of the network was tested in several ways, both in the presence and in the absence of the HC. The ability of the network to recall the stored memories was assessed by presenting parts of these patterns as inputs to see whether they can be completed. Partial patterns consisted of the correct pattern in two input areas (say, $\mathbf{x}^{A_1}\mathbf{x}^{B_1}$), and random activation in the third area ($\mathbf{x}^{C}$). Next, the activities in E/P were computed. In cases when the HC was present, the E/P pattern was then compared to each of the stored representations, and, if it was sufficiently close to one of them (using an arbitrary threshold of 20 on the squared distance), it was replaced by that stored pattern; otherwise, it was left unchanged. This is intended to correspond to familiarity-based modulation of hippocampal processing, and allows good separation of E/P patterns corresponding to partial cues from the representations of other possible input patterns. It also allows selection of the correct stored pattern, as evidenced by Figure 2a. Then the cortical network was allowed to run for 20 iterations (with $\mathbf{x}^{A_1}\mathbf{x}^{B_1}$ clamped). Some examples of the convergence to the correct pattern in area C (after consolidation, with and without HC) are shown in Figure 2b. The final activation patterns in area C were classified as follows: correct pattern ($\mathbf{x}^{C_1}$), other memorized pattern (one of $\mathbf{x}^{C_2}$–$\mathbf{x}^{C_8}$), other (not memorized) valid pattern ($\mathbf{x}^{C_9}, \mathbf{x}^{C_{10}}$), and none of the above. Small errors in the recalled patterns (a square distance of 2) were allowed in assessing the first three classes. The frequency distribution of errors (squared distance from the correct pattern) was also determined.

The main results of the simulations are displayed in Figure 3. Before memoriz-

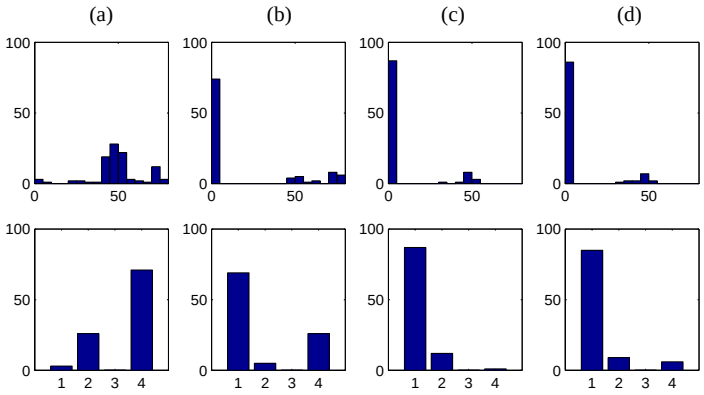

Figure 3: Recall performance. The four columns of plots represent the four cases in which performance was tested. These are: (a) after generic pretraining, either before memorizing the specific patterns, or after memorizing them but with the hippocampus removed before consolidation could take place; (b) after memorizing the episodes, in the presence of HC; (c) after consolidation, in the presence, or (d) in the absence of HC. The top row of plots shows the frequency distribution of the distance between the pattern recalled (after 20 iterations) and the correct pattern; the bottom row shows the relative frequencies of the qualitatively different possible results of the recall process: (1) correct pattern; (2) other memorized pattern; (3) other valid pattern; (4) none of the above.

ing the episodic patterns, or, equivalently, if the hippocampus is removed before any consolidation can occur, only semantic knowledge about the input patterns is available. Consequently, given the partial pattern $\mathbf{x}^{A_1}\mathbf{x}^{B_1}$, all valid patterns in area C ($\mathbf{x}^{C_1}$–$\mathbf{x}^{C_{10}}$) are still equally likely, and are ultimately generated in approximately equal proportions by the network. However, the network often does not settle into one of the valid patterns within the first 20 iterations. This situation changes dramatically after the selected patterns are stored by the hippocampus. The E/P representations corresponding to the partial patterns are now recognized and completed by the hippocampus in most cases, which leads to good completion in the input areas due to the existing generative model relating activity in $\mathbf{y}$ to $\mathbf{x}^C$.

The subsequent consolidation process alters the cortical weights so that the cortical network now represents a different probability distribution over the inputs; in particular, the probabilities corresponding to the memorized patterns are increased relative to all other patterns. This results in some further improvement in the pattern completion performance of the full network (neocortex + hippocampus). More importantly, however, the changed cortical weights can support the recall of the stored patterns on their own, so that the removal of the hippocampus at this stage hardly affects the recall of consolidated patterns.

An important point about the consolidation process is that it should not impair the ability of the network to represent valid input patterns other than the ones memorized and consolidated, or to reconstruct these other input patterns from their E/P representations. It is relatively easy to come up with a learning algorithm which completes *any* input pattern to one of the patterns stored, but, among other things, such a network would find it difficult to memorize any additional input patterns subsequently. Our model, on the other hand, retains its ability to represent arbitrary combinations of the valid input patterns after consolidation (and, indeed, at all stages of training). We checked this by presenting input patterns different from the memorized episodes, and verifying that the one-step reconstruction in the input areas (after processing in E/P and, if applicable, the HC) was closer to the

pattern presented than to any other valid pattern. We found this to be the case for all the patterns we tested, at all stages of training (after pre-training), both with and without the HC.

4 Discussion

By placing consolidation in the context of a hierarchy of cortical areas, we have been able to consider a number of crucial issues that simpler models find hard to address, in particular the way that storing episodic information might be temporarily parasitic on the (prior) storage of semantic information. Thus, it also captures and refines prior qualitative insights such as that the hippocampus might somehow store *pointers* to cortical memories (that, in this case, are dereferenced via the population code in E/P).

Various extensions to the model are desirable. Two straightforward extensions are to a deeper cortical hierarchy (possible because the RBM model has a natural hierarchical extension) and to slow, hippocampally-independent cortical learning of specific information (via an explicit gross change to input statistics that is normally created implicitly by the hippocampus during consolidation). We also intend to model the hippocampal system in much more detail, particularly in the context of fast and slow learning. The longer-term stability of both hippocampal and cortical representations also needs to be addressed, together with the phenomena of normal forgetting.

References

[1] Squire, L.R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys and humans. Psychol. Rev., 99, 195-231.

[2] Alvarez, P. and Squire, L.R. (1994). Memory consolidation and the medial temporal lobe: A simple network model. Proc. Natl. Acad. Sci. USA, 91, 7041-45.

[3] McClelland, J.L., McNaughton, B.L., and O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psych. Rev., 102, 419-457.

[4] Murre, J.M. (1997). Implicit and explicit memory in amnesia: some explanations and predictions by the TraceLink model. Memory, 5, 213-32.

[5] Nadel, L. and Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. Curr Opin Neurobiol., 7, 217-27.

[6] Hinton, G. and Sejnowski, T.J. (eds.) (1999). Unsupervised learning. Cambridge, MA, MIT Press.

[7] Freund, Y. and Haussler, D. (1992). Unsupervised Learning of Distributions of Binary Vectors Using 2-Layer Networks. In Moody, J.E., Hanson, S.J., and Lippmann, R.P. (eds.), Advances in Neural Information Processing Systems 4. San Mateo, CA, Morgan Kaufmann.

[8] Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA, 79, 2554-8.

[9] Hinton, G. and Sejnowski, T.J. (1986). Learning and relearning in Boltzmann machines. In Rumelhart, D.E. and McClelland, J.L. (eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations. Cambridge, MA, MIT Press.

[10] Hinton, G. (2000). Training Products of Experts by Minimizing Contrastive Divergence. Technical Report GCNU 2000-004.

[11] Wilson, M.A. and McNaughton, B.L. (1994). Reactivation of hippocampal ensemble memories during sleep. Science, 265, 676-679.