Hippocampally-Dependent Consolidation in a Hierarchical Model of Neocortex

Szabolcs Káli[1,2]   **Peter Dayan[1]**

[1]Gatsby Computational Neuroscience Unit
University College London
17 Queen Square, London, England, WC1N 3AR.

[2]Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139, U.S.A.

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 (from several days in rats 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 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.