Sleep and memory

Recent experimental and theoretical advances suggest that memories may be reorganized in the cortex during sleep.

During sleep, our brains are highly active. The low-amplitude, high-frequency activity in the neocortex characteristic of the awake state is replaced with high amplitude, low-frequency rhythms during slow-wave sleep [1]. It would seem unlikely that the extensive cortical activity during sleep does not have some purpose; however, there is still no consensus on why we need to sleep. One intriguing possibility is that information acquired during the day is compared during sleep with older memories [2]. Previous neural network models included such a 'sleep phase' to calibrate the storage of memories acquired by Hebbian mechanisms [3–5]. Recent recordings from the hippocampus [6], and a new neural network model [7], lend experimental support and computational motivation to the possibility that we may sleep in order to organize efficient cortical representations of experience.

Cortical representations of objects and events are widely distributed in the cerebral cortex. Thus, the representation of a violin might be stored in areas as diverse as the visual cortex, for its shape, the auditory cortex, for its sound, the parietal cortex, for how it may be grasped, and the motor cortex, for how it is played [8]. Problems arise when new experiences and objects must be integrated with existing information that is widely distributed. Learning algorithms designed for artificial neural networks that use such distributed representations can suffer from 'catastrophic interference' when new information is stored in the same neural circuits as old information [9]. Therefore, the brain must solve two problems during learning: where to make the changes needed to create a new memory; and how to make changes that are compatible with previously stored memories.

There appears to be a period of consolidation before a memory becomes permanently stored. Thus, lesions of hippocampal formation, including the parahippocampal, perirhinal and entorhinal cortices, lead to memory deficits for up to 6 weeks following learning in monkeys [10], and more than a year in man. After this period of consolidation, lesions of the same areas are less disruptive, implying that the memories are stored elsewhere. Until recently, the processes that may occur in the cortex during the period of consolidation could only be inferred indirectly from such lesion experiments.

Wilson and McNaughton [6] have reported changes that occur in the correlations between hippocampal neurons as a consequence of a new learning experience. They were able to record simultaneously from over 40 neurons in the hippocampus before and after a rat explored an environment. Many hippocampal neurons respond to places in the environment [11]. In these experiments, the activities of neurons that had neighboring place fields and fired together during exploration of an environment became more highly correlated during sleep in comparison with their activities during preceding sleep episodes. The correlated firing of neurons in the hippocampus, reflecting newly acquired experience, may be 'played back' to the neocortex through feedback projections.

It has been suggested that the purpose of such playback from the hippocampal formation is to provide a 'teaching' signal for the neocortex, which would receive a summary of events that have been stored temporarily in a raw form in the hippocampal formation [9,12]. Thus, the neocortex during the awake state provides the hippocampal formation with a detailed description of the day's events; during sleep, the hippocampus plays back some version of these events to the neocortex, where permanent memory representations are gradually formed over repeated episodes of sleep. Why would the brain go to so much trouble to reorganize cortical memories?

Hinton *et al.* [7] have provided an elegant new theoretical framework for creating efficient memory representations in hierarchical neural network models. In this model (Fig. 1), the feedback connections generate patterns on the input layers of the network that correspond to the representations at the higher level, when the external inputs to the cortex and feedforward processing have been suppressed. During this generative sleep stage, the strengths of the feedforward synaptic strengths are altered. Conversely, during the awake stage, the feedback connections are suppressed and the sensory inputs drive the feedforward system, during which the weights on the feedback connections can be altered.

The synaptic learning rule used in both phases is a local correlational, or Hebb rule, that reduces the mismatch between the feedforward-driven activity and the feedback-driven reconstruction. This two-phase process produces internal, hierarchical representations of experiences that are economical and able to generate typical input patterns. The learning mechanisms needed are biologically possible as, unlike previous learning algorithms that required both a 'teacher' to provide a comparison between the desired and actual outputs of the network and 'backpropagation' of the resulting error signal through the network, the wake–sleep model only depends on locally available signals and there is no teacher.

The wake–sleep learning algorithm attempts to capture the statistics of sensory inputs with an internal code that is capable of representing component features that are common to many objects. Because these statistical components are not apparent without comparing many sensory experiences, the training process is gradual, in the sense that only small changes are made during any one wake–sleep cycle. Another way to view the hippocampal representation, according to this view, is as a repository of previous items and events which are played back into the neocortex until the information that they contain is adequately summarized by the neocortical representation along with many similar items and events [9,13,14].

The wake–sleep model has several virtues. It is consistent with the recent data on memory consolidation and hippocampal playback. It also makes testable predictions for the function of feedback connections and the conditions when plasticity should occur. It has proved difficult to assign any function for cortical feedback connections and the wake–sleep model explains why. Experiments designed to test the effects of these connections have been performed during the awake state, but according to the model, they should only drive physiological activity during sleep. Also, the model predicts that the feedback and feedforward projections should be modifiable at different times: feedforward connections during sleep, and feedback connections during the awake state. The types of experiment that are now possible using multielectrode recordings should allow these predictions to be tested directly [6].

The wake–sleep model needs further refinement. In the cortex, the feedforward and feedback systems involve two different sets of neurons, but there is only one in the wake–sleep model. This separation may make it easier to regulate the activity in these pathways independently. There are two sleep states, slow-wave sleep and rapid eye movement (REM) sleep [2], and two awake states in rats, a state of low frequency theta in the hippocampus during exploration, and a state that resembles the fast activity characteristic of slow-wave sleep during rest. I suggest that both phases of learning can occur during both wake and sleep, one phase during one of their two states, the other phase during the other state.

Let us call the 'sleep' phase of the wake–sleep model the 's-phase' and the 'wake' phase the 'w-phase'. Then, during true sleep, REM might correspond to the w-phase of learning. During REM sleep, the visual cortex is driven by brain-stem activity and the hippocampus exhibits theta rhythm. Slow-wave sleep would correspond to the s-phase of learning, driven by high-frequency activity from the hippocampal formation. During the true awake state, the w-phase of learning would correspond to exploration, and the s-phase of learning would occur during rest when the hippocampus is driving feedback connections. In a sense, the brain during this awake s-phase is 'day dreaming'. According to this suggestion, the interplay of the s-phase and w-phase may occur on different

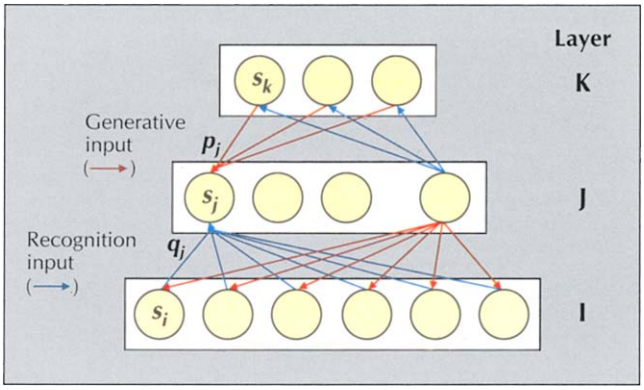

Fig. 1. Wake–sleep network model. Sensory inputs originate in the bottom layer and feedforward connections carry this information through a layer of hidden units to the top layer in the recognition or wake phase. The feedback connections from top to bottom provide a generative input to the bottom layer during the sleep phase. s_j indicates unit j in layer J; p_j is its firing probability when driven by the feedback, generative connections; and q_j is its firing probability when driven by the feedforward, recognition connections. (Adapted from [7].)

time scales, ranging from minutes during periods of alertness to hours during sleep, perhaps allowing learning and consolidation to occur using different molecular mechanisms.

The wake–sleep model is also limited to a passive, unsupervised form of learning that is entirely driven by the statistics of sensory states. Not all sensory inputs are equally important, and some tasks might require special representations. It would be easy to add an attentional mechanism that would modulate the learning rate according to the significance of the stimulus. There could also be biases in cortical representations at birth that are specified during development, which could incorporate a prior probability distribution representative of the real world. The goal-directed reinforcement learning system requires rewards and penalties and involves subcortical as well as cortical structures. Unsupervised wake–sleep learning and other forms of learning could work together, biasing, shifting and adapting cortical representations to ensure survival in complex and uncertain environments.

The recent experimental and theoretical advances on sleep and learning are exciting because they suggest a possible resolution to one of the greatest mysteries in biology, the nature and function of sleep. The results so far are incomplete and tentative, but they should lead us toward further advances that will widen our understanding of the "Sleep that knits up the ravell'd sleave of care".

References

1. Steriade M, McCormick DA, Sejnowski TJ: **Thalamocortical oscillations in the sleeping and aroused brain.** *Science* 1993, **262:**679–685.
2. Hobson JA: *The dreaming brain.* New York: Basic Books; 1988.
3. Crick F, Mitchison G: **The function of dream sleep.** *Nature* 1983, **304:**111–114.
4. Hopfield JJ, Feinstein DI, Palmer RG: **'Unlearning' has a stabilizing effect in collective memories.** *Nature* 1983, **304:**158-159.
5. Ackley DH, Hinton GE, Sejnowski TJ: **A learning algorithm for Boltzmann Machines.** *Cogn Sci* 1985, **9:**147–169.

6. Wilson MA, McNaughton BL: **Reactivation of hippocampal ensemble memories during sleep.** *Science* 1994, **265**:676–679.
7. Hinton GE, Dayan P, Frey BJ, Neal RM: **The 'wake–sleep' algorithm for unsupervised neural networks.** *Science* 1995, **268**:1158–1161.
8. Damasio AR, Tranel D: **Nouns and verbs are retrieved with differently distributed neural systems.** *Proc Natl Acad Sci USA* 1993, **90**: 4957–4966.
9. McClelland JL, McNaughton BL, O'Reilly RC: **Why there are complementary learning systems in the hippocampus and neocortex: insights from the success and failures of connectionist models of learning and memory.** *Psychol Rev* 1995, in press.
10. Zola-Morgan SM, Squire LR: **The primate hippocampal formation: evidence for a time-limited role in memory storage.** *Science* 1990, **250**:288–290.
11. O'Keefe J, Nadel L: *The Hippocampus as a Cognitive Map.* Oxford: Clarendon Press; 1978.
12. Buzsaki G: **Two-stage model of memory trace formation: a role for 'noisy' brain.** *Neuroscience* 1989, **31**:551–570.
13. Marr D: **Simple memory: A theory for the archicortex.** *Phil Trans Roy Soc Lond [Biol]* 1971, **262**:23–81.
14. Gluck M, Myers CE: **Hippocampal mediation of stimulus representation: a computational theory.** *Hippocampus* 1993, **3**: 491–516.

Terrence J. Sejnowski, Howard Hughes Medical Institute, Salk Institute for Biological Studies, La Jolla, California 92037, USA, and Department of Biology, University of California at San Diego, La Jolla California 92093, USA.