French, R. M., Ans, B., & Rousset, S. (2001). Pseudopatterns and dual-network memory models: Advantages and shortcomings. In ***Connectionist Models of Learning, Development and Evolution.*** (R. French & J. Sougné, eds.). London: Springer, 13-22.

Pseudopatterns and dual-network memory models: Advantages and shortcomings

Robert M. French, Bernard Ans, & Stéphane Rousset

Abstract

The dual-network memory model is designed to be a neurobiologically plausible manner of avoiding catastrophic interference. We discuss a number of advantages of this model and potential clues that the model has provided in the areas of memory consolidation, category-specific deficits, anterograde and retrograde amnesia. We discuss a surprising result about how this class of models handles episodic ("snap-shot") memory — namely, that they seem to be able to handle both episodic and abstract memory — and discuss two other promising areas of research involving these models.

1. Introduction

Neural networks typically store patterns in a single set of weights. This lack of modularity can mean that when new patterns are learned by the network, the new information may radically interfere with previously stored patterns. This problem, called catastrophic interference, was first brought to light by McCloskey and Cohen [12] and Ratcliff [15] and has been studied by numerous authors since (see [8] for a review). In this paper we will discuss a particular connectionist architecture designed to overcome this problem: the "dual-network" architecture. This model is loosely patterned after the "hippocampal-neocortical" architecture of the brain. In cognitive tasks, such as recall, pattern recognition, etc., humans do not experience catastrophic interference. McClelland, McNaughton and O'Reilly [11] suggested that the reason for this is the brain's bi-partite hippocampal-neocortical division of labor. New patterns are initially learned only by the hippocampus. The hippocampus then slowly trains the neocortex and, in this way, new patterns to do not interfere with already stored patterns.

There are a number of problems with this account; in particular, it does not explain why the transferred hippocampal patterns — regardless of the speed with they are learned by the neocortex — do not interfere with the patterns already in neocortex. Whether the new information "trickles" into the neocortex or is learned quickly is not the key issue here; new information, whether from the environment or

from the hippocampus, can still overwrite old information in neocortex. French [6] and Ans & Rousset [1, 2] independently developed dual-network models that overcame this problem based on the use of pseudopatterns [17] to transfer information from one network to the other.

The key to these the dual-network models is pseudopattern transfer. In this paper we will discuss this crucial information transfer mechanism. We suggest that a clearer understanding of pseudopattern information transfer may be able to provide insights into the function of REM sleep, memory consolidation, category-specific deficits, anterograde and retrograde amnesia. We will discuss a number of potential problems with this type of mechanism and will suggest how the pseudopattern generation mechanism might be optimized.

2. Sensitivity and stability to new information

One of the most important problems facing connectionist models of memory — in fact, facing *any* model of memory — is how to make them simultaneously sensitive to, but not disrupted by, new input. One solution to this problem — arguably, the solution discovered by evolution for the human brain — is to have two separate storage areas: one for new information (the hippocampus), the other for previously learned patterns (the neocortex). The idea would be that the hippocampus gradually transfers the new patterns to the neocortex and, in this way, catastrophic interference would be avoided [11]. The major problem with this suggestion is that the rate, however slow or fast, at which new information is learned by the neocortex does not prevent the overwriting of previously learned patterns by the new patterns. The key to overcoming catastrophic interference is to *interleave* previously learned patterns (or some approximation of these patterns) with the new patterns during learning.

3. Dual-network memory models

To overcome the problem of catastrophic interference French [6] and Ans & Rousset [1, 2] independently proposed dual-network memory models. Even though their respective models differ in a number of respects, the essence of the two architectures is largely the same. The overall system consists of two separate, coupled networks. New information arrives at only one of the networks and is interleaved with information from the other network. The newly learned information is then passed from the first network to the second. Information is transferred from one network to the other by *pseudopatterns* [17].

Consider the following problem: A neural network has learned a number of input-output patterns corresponding to some underlying function *f*. We have no access to the network's weights, its connection topology, etc, and, most importantly, *the original patterns learned by the network are no longer available*. How can we, nonetheless, get an approximation of the original function *f*?

One solution to this problem consists of sending input (in the simplest case, random input) into the network and observing the output for each random input. We thus create a series of *pseudopatterns*, ψ_i, where each pattern ψ_i is defined by a random input and the output of the network after that input has been sent through it.

This set of pseudopatterns reflects the originally learned function *f*. The greater the number of pseudopatterns, the better the approximation that can be obtained of the originally learned function *f*. Pseudopatterns were first introduced by Robins [17] to overcome catastrophic interference. Robins suggested that when a network had to learn a new pattern, a number of pseudopatterns be generated. Then, instead of learning just the new pattern, *P*, the network would be trained on the new pattern plus the set of pseudopatterns that reflected what it had previously learned. In this way, the new pattern would be interleaved with patterns that, even though they were not the originally learned patterns, nonetheless reflected the original function learned. Robins showed that his technique did, indeed, reduce catastrophic interference [4, 16, 17, etc.].

This technique was used successfully by French [6] and Ans & Rousset [1] as the main mechanism of information transfer between two separate networks. The idea is to have two networks (for simplicity we will call these two networks the "hippocampal" and the "neocortical" networks, although these designations should not be taken too literally) that exchange information by means of pseudopatterns (the neural correlate of pseudopattern information transfer might be experienced as dream sleep [18]). When a new pattern, *P*, is to be learned by the hippocampal network (new information from the environment is learned exclusively by the hippocampal network), there are (at least) three ways to proceed:

i) A set of neocortical pseudopatterns $\{\psi_i\}^{NEOCORTEX}$ is created. Hippocampal network learning continues until the network has learned the new pattern as well as the neocortical pseudopatterns to criterion [6].

ii) A set of neocortical pseudopatterns $\{\psi_i\}^{NEOCORTEX}$ is created. Hippocampal network learning continues only until the new pattern *P* falls below criterion, even though the network may or may not have learned the neocortical pseudopatterns to criterion [17].

iii) There is no fixed set of neocortical pseudopatterns. Rather, new neocortical pseudopatterns are continually generated while the hippocampal network gradually learns the new pattern, *P*. Hippocampal network learning continues until *P* falls below criterion [1, 2].

In Ans & Rousset's model the inputs used to produce the pseudopatterns are not simply random patterns, as they are in the other two models, but rather they result from a "reverberation" between the first and second layer of the network in their network of origin. The first and second layers of both networks in the Ans & Rousset model constitute an autoassociator. When a pseudo-input is given to the network, it is cycled from the first layer to the second to the first to the second, etc., until a maximum number of cycles has been reached. This "input attractor" (that has not necessarily reached a stable state) is then fed through the network to produce the pseudopattern that will be learned by the other network. Pseudopatterns are also used to transfer the newly learned information from the hippocampal network to the neocortical network. To do this, pseudopatterns $\{\psi_i\}^{HIPPOCAMPUS}$ are generated by the hippocampal network and are then learned by the neocortical network.

4. Advantages of the dual-network approach

It has been shown [1, 2, 6] that this type of approach does, indeed, eliminate catastrophic forgetting. The forgetting curves for this type of model are far more realistic than in any standard backpropagation or Hopfield network. For example, in one experiment French [6] compared a standard backpropagation network to a dual-network memory on a task that consisted of having the networks sequentially learn a total of 20 items. After learning the 20th item, the error performance is measured for each item in the 20-item list. For standard backpropagation, all 19 previous items were well above the 0.2 error threshold, whereas for the dual-network memory, forgetting was far more gradual (e.g., 8 of the most recently learned items were still at or below the 0.2 threshold). Ans & Rousset [2] have shown similar improvements in forgetting with their dual-network reverberating architecture.

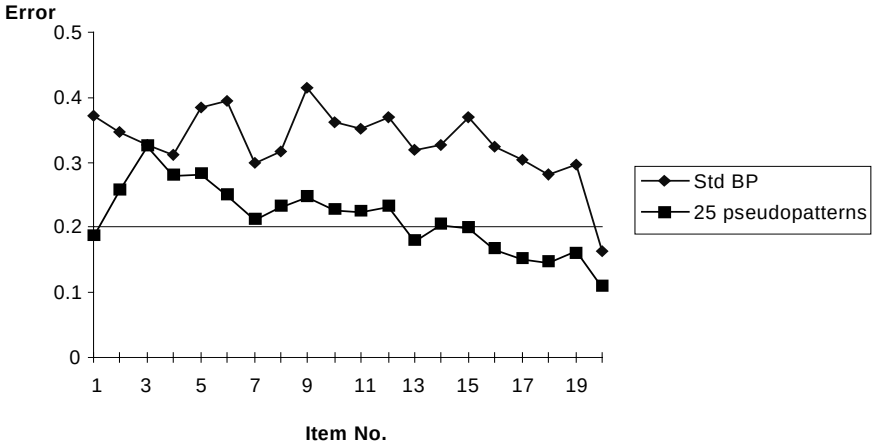

Figure 1: Amount of error for each of 20 items learned sequentially after the final item had been learned to a 0.2 criterion [6].

It also seems that this process of passing information back and forth between two networks using pseudopatterns may lead to representational compression over time, which may help explain rather puzzling category-specific losses observed in certain cases of anomia [5, 7].

Further, Robins & McCallum [16] suggest that this pseudopattern information transfer in this type of dual-network system is the primary means of long-term memory consolidation. These authors claim that memory consolidation by pseudopattern information transfer to the neocortex is "a computational model of sleep consolidation" that is supported by "psychological, evolutionary and neurophysiological data (in particular accounting for the role of the hippocampus in consolidation)." (For a review, see [20]). Related conclusions concerning "off-line" mutual reactivation of memory traces in both hippocampus and neocortex are discussed in [21]. Even though the contention that memory consolidation is contested by certain authors (e.g., [24]), the point remains that a mechanism resembling pseudopattern information transfer may well be responsible, at least in part, for LTM memory consolidation.

Ans & Rousset [2] have also argued that lesioning the flow of pseudopattern transfer from the “hippocampal” to the “neocortical” network should induce an anterograde amnesia behavior and, because pseudopatterns from the neocortical network would continue to refresh the hippocampal network, no severe retrograde amnesia be observed. In the other direction, damage to neocortical to hippocampal transfer would mean that pre-lesion information would continue to be consolidated, but there would be post-lesion catastrophic interference by new patterns of information already present in hippocampus.

5. Potential difficulties

Potential difficulties with this architecture fall broadly into three categories.

Episodic (“snap-shot”) memory

It is well known that people can recall “snap-shot” episodes of experience, i.e., precise events that occurred at a specific time and a specific place. These so-called snap-shot memories are widely believed to be stored in the hippocampus, although undoubtedly prefrontal cortex also plays some lesser role in episodic memory [3, 19]. A common concern that has been voiced about the pseudopattern transfer mechanism involves its ability to preserve these snap-shot memories. This concern arises from the intuition that the random nature of pseudopattern input would tend to “blur to abstraction” the originally learned patterns. In other words, precise snapshot memories corresponding to particular patterns would be lost.

This important concern merits closer examination. In Ans & Rousset's dual-network model [1], the pseudopattern inputs are first reverberated between the input and hidden layers towards an attractor. Since the input-hidden layers of each network act much like the auto-associative "clean-up" units used in many current connectionist models, intuitively, this would seem to lead to one of the following situations:

- If the input attractor basins were very large, there might be a problem of coverage of the input space, i.e., even though in this case pseudopatterns would indeed tend to be the originally learned items (thereby satisfying the snap-shot memory criterion), those with small attractor basins could be systematically neglected for items with much larger attractor basins. This would mean that the neglected items would never be transmitted from one network to the other and would therefore be forgotten by the network;
- If the input attractor basins were relatively small, then many pseudo-inputs would not fall into attractor basins corresponding to previously learned input and the originally learned snap-shot memories would be lost.

In one of the most surprising results to come out of dual-network memory model research, it would seem that it is possible that neither of these situations occurs. The dual-network model would seem to provide the best of both worlds: abstraction occurs but, at the same time, specific memories are preserved.

In a preliminary study, Ans & Rousset [2] trained the hippocampal network on a set of 20 arbitrary input-output patterns, $\{P_i\}_{i=1}^{20}$. Then, in order to transfer this learning to the neocortical network, they produced 32-bit random pseudo-inputs and reverberated these inputs towards attractors in the hippocampal network in order to produce the hippocampal pseudopatterns that would subsequently be learned by the neocortical network. Each 32-bit pseudo-input was compared to the inputs of the 20 previously learned patterns P_i. A normalized Euclidean distance metric was defined on the input space as follows: $d(X,Y) = \left[\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2\right]^{1/2}$, where N is the length of the input vector (in this case, 32) for each pattern. This metric keeps distance between any two input patterns between 0 and 1. Any pseudo-input that was within 0.5 of the input of any previously learned pattern, P_i, *was rejected.* This draconian filtering meant that only 13% of the total number of pseudo-inputs generated were actually used to train the neocortical network. In short, the neocortical network was trained only on patterns that specifically *did not resemble* the patterns previously learned by the hippocampal network. One might expect, because of the artificially forced dissimilarity between the hippocampal pseudopatterns and the patterns originally learned by the hippocampal network, that the original patterns, $\{P_i\}_{i=1}^{20}$, would not be accurately transferred to the neocortical network. Surprisingly, this proved not to be the case. In fact, all twenty of the originally learned patterns were successfully transferred to neocortical memory! This result has been tested numerous times and seems to be robust.

Although considerable research still needs to be done on this aspect of dual-network memory models, this result would suggest that this type of memory model not only accommodates the generally accepted idea that hippocampal memory is appropriately modeled by a sparse auto-associative network [13, 14, 23, etc.], thereby allowing it to precisely recover previously learned information, but simultaneously allows abstraction to take place in the neocortical network.

Contextualizing pseudopattern generation

When we encounter a new pattern — say, the first time we see a yo-yo — we must be able to learn this pattern without it interfering with other patterns we have already learned. In the dual-network model this is achieved by interleaving the new pattern with pseudopatterns that reflect previously learned patterns. But there is a problem with this — namely, that pseudopatterns reflecting *the undifferentiated contents* of neocortical memory are interleaved with each new pattern that must be learned in the hippocampal memory. This is surely not necessary to prevent catastrophic interference and, moreover, it is ludicrous to suppose that we need an approximate copy in hippocampus of everything that has ever been stored in long-term memory every time we learn a single new association.

In particular, in the case of our first encounter with a yo-yo, we can reasonably assume that our internal representations of patterns relating to sabre-tooth tigers and unicorns, computer memory chips, and the Crab Nebula will almost certainly be

orthogonal to the representation we will develop of "yo-yo." This would mean that the internal representation for the new pattern "yo-yo" would not interfere with those representations for unicorns, computer chips and clusters of stars. But it certainly *could* interfere with our representation for concepts like "wheel," "pendulum", etc.

What needs to happen is that the input from the "yo-yo" pattern must activate similar concepts in long-term memory and that *these related concepts* need to be refreshed so as not to be overwritten. But how can this been done in a reasonable manner? Assume that the pattern to be learned by the hippocampal network is $P: i \rightarrow o$. One suggestion might be to give i as input to the neocortical network in order to produce a similar pseudopattern. This would, however, produce a neocortical pseudopattern ψ: $i \rightarrow \hat{o}$. The problem is that the hippocampal network (or any network) *cannot* simultaneously learn both P and ψ.

In fact, Ans & Rousset have shown that when neocortical pseudopatterns whose input resembles the input of the pattern to be learned are interleaved with the new pattern, the network frequently fails to converge. Performance is considerably better with pseudopatterns whose initial input (i.e., before reverberation) is random.

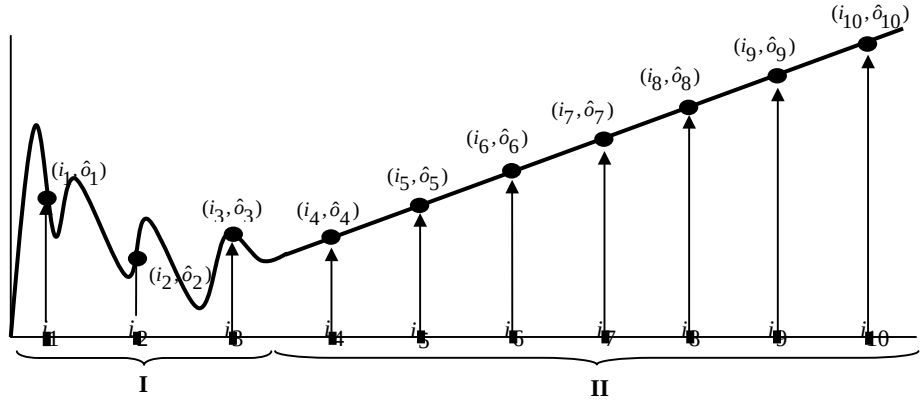

Figure 2: A uniform distribution of pseudo-inputs to generate pseudo-patterns leads to an over-exploration of Region II and an under-exploration of Region I.

Optimizing pseudopattern generation

The final issue that we will consider in this paper is closely related to the previous problem of pseudopattern contextualization. The question is: Can pseudopattern generation be optimized to improve recovery of the originally learned function and is the brain doing anything similar to this kind of optimization?

We begin this discussion with a problem first posed by John Holland [10]. Suppose we have a *two*-armed bandit which has two payoff arms. One arm pays off with a ratio of p, the other with a ratio of q where $p > q$, but we do not know which arm gives which payoff. We have N tokens and we wish to maximize our earnings. If we knew which arm was which, we would, of course, put all of our tokens in the arm with payoff ratio p. But we don't have this information, so we must "waste" some of our supply of tokens to try to determine which arm pays off more. One strategy might be to decide to allocate N/4 tokens to the first arm, N/4 tokens to the second and, then, whichever arm had produced the greatest payoff, put the remaining N/2

tokens in the slot corresponding to that arm. The problem with this strategy, of course, is that if the p ratio is 1000 and q is 1, then we would realize almost immediately which is the better arm and most of the N/4 tokens put in the arm corresponding to payoff q would be wasted.

Holland proved a theorem concerning the strategy to adopt. Basically, "the loss rate will be optimally reduced if the number of trials allocated [to the apparently most promising arm] grows slightly faster than an exponential function of the trials allocated [to the apparently least promising] arm." ([10], p. 83).

A related problem arises in pseudopattern generation. To see why, consider the following function that has been learned by a neural network. We wish to recover this function as best as possible using pseudopatterns.

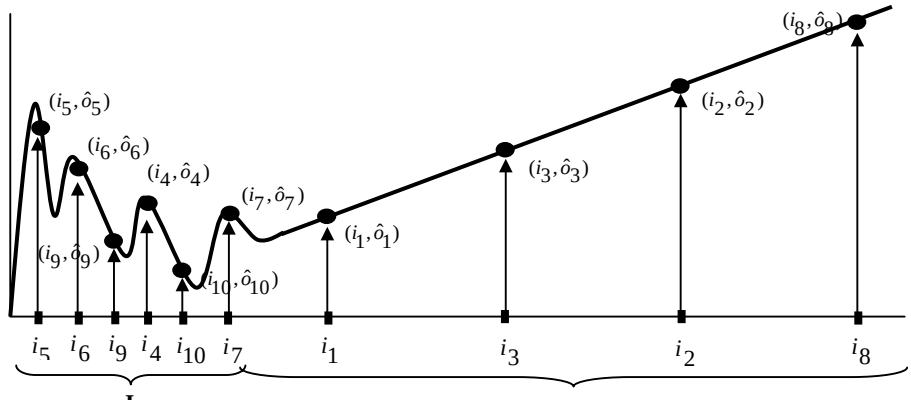

Figure 3: A better exploration of the space is obtained by using feedback from the patterns as they are generated. In this way, more of the patterns are concentrated in the more complex region of the function.

The set of pseudopatterns in Figure 4, $\{\psi_k : i_k \rightarrow \hat{o}_k\}_{k=1}^{10}$ does, indeed, approximate the originally learned function, but could we have done better by a more judicious choice of inputs to probe the network? Clearly, the answer is yes. In Figure 2, seven of the ten pseudopatterns fall in Region II, which does not need to be explored as carefully as Region I. Notice that nowhere have we used the information provided by the first pseudopatterns to determine the input to create the later ones. In a manner similar to the strategy for optimally allocating our tokens in the two-armed bandit problem, once we begin to be certain that we have a good representation of part of the function (Region II), we concentrate our remaining pseudo-inputs on other less well-understood regions (Region I).

If, for example, the network makes a linear interpolation prediction as to where the output from a given pseudo-input should fall, and this turns out to be correct (or close to correct), then this tells the system that it is probably in a well-understood area. In other words, in the first case of pseudo-input selection (Figure 2), we are not using the output of the network to modify subsequent pseudo-input selection. In the second case (Figure 3), as soon as the system begins to be able to accurately predict the outputs for a particular region of the function, then we can devote more of our pseudo-input "energy" to other regions of the function. This does not, of course,

mean that we cease further exploration of the well-understood region. The function could be highly varying, in that region and we were simply lucky in our prediction. But it does mean that we devote exponentially less of our pseudo-inputs to that area as our information of what the function in that region is likely to be improves.

If REM sleep and pseudopattern generation are related, we can probably conclude that pseudopattern generation is not random. It would seem likely that some kind of pseudopattern optimization is being carried on by the brain. What and how this works is an important question for further study. We suggest that perhaps there is some stochastic "2-armed bandit" optimization method being used by the brain in the type of noise that it "selects" to send through the system in order to improve the consolidation of new information in neocortex.

6. Conclusions

We have attempted to show that, while the dual-network memory model has great potential for explaining a number of phenomena related, in particular, to memory consolidation, there are still a number of problems that need to be considered. We have considered three of these problems relating to episodic memory, contextualization of pseudopattern generation and pseudopattern optimization. We have attempted to show why these areas pose problems for the dual-network model and have suggested ways in which these problems might be able to be overcome.

Acknowledgments

This research was supported in part by a research grant from the European Commission (HPRN-CT-1999-00065) and the Belgian government (IUPA P4/12).

References

1. Ans, B., & Rousset, S. (1997). Avoiding catastrophic forgetting by coupling two reverberating neural networks. *Academie des Sciences, Sciences de la vie, 320*, 989-997.
2. Ans, B., & Rousset, S. (2000). Neural Networks with a Self-Refreshing Memory : Knowledge Transfer in Sequential Learning Tasks without Catastrophic Forgetting. *Connection Sciences, 12*, 1, *1-19*
3. Cohen, J. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the Hippocampal System*. Cambridge: MIT Press.
4. Frean, M., & Robins, A. (1998). Catastrophic forgetting and "pseudorehearsal" in linear networks. In Downs T, Frean M., & Gallagher M (Eds.) *Proc. of the 9th Australian Conference on Neural Networks*, 173-178, Brisbane: U. of Queensland
5. French, R. M., & Mareschal, D. (1998). Could Category-Specific Semantic Deficits Reflect Differences in the Distributions of Features Within a Unified Semantic Memory? In *Proceedings of the Twentieth Annual Cognitive Science Society Conference*. NJ:LEA. 374-379.
6. French, R. M. (1997a). Pseudo-recurrent connectionist networks: An approach to the "sensitivity–stability" dilemma. *Connection Science, 9*(4), 353-379.

7. French, R. M. (1997b). Selective memory loss in aphasics: An insight from pseudo-recurrent connectionist networks. In J. Bullinaria, G. Houghton, D. Glasspool (eds.). *Connectionist Representations: Proceedings of the Fourth Neural Computation and Psychology Workshop*. Springer-Verlag. 183-195.
8. French, R. M. (1999). Catastrophic Forgetting in Connectionist Networks. *Trends in Cognitive Sciences, 3*(4), 128-135.
9. Hebb, D. O. (1949). *Organization of Behavior*. New York, N. Y.: Wiley & Sons.
10. Holland, J. (1975). *Adaptation in natural and artificial systems.* Ann Arbor, MI: The University of Michigan Press.
11. McClelland, J., McNaughton, B., & O'Reilly, R. (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. *Psychological Review*. *102*, 419–457.
12. McCloskey, M., & Cohen, N. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, *24*, 109-165.
13. McNaughton, B., & Morris, R. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences, 10*, 408-415.
14. McNaughton, B., & Nadel, L. (1990). Hebb-Marr networks and the neurobiological representation of action in space. In M.A. Gluck, & D. Rumelhart (Eds.) *Neuroscience and Connectionist Theory*. Hillsdale, NJ: LEA, 1-63.
15. Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97*, 285-308
16. Robins, A., & McCallum, S. (1998). Pseudorehearsal and the catastrophic forgetting solution in Hopfield type networks. *Connection Science, 10*, 121 - 135
17. Robins, A. (1995). Catastrophic forgetting, rehearsal, and pseudorehearsal. *Connection Science, 7*, 123 - 146.
18. Robins, A. (1996). Consolidation in neural networks and in the sleeping brain. *Connection Science, 8*, 259 - 275
19. Squire, L. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99*, 195-231.
20. Stickgold, R. (1999). Sleep: off-line memory reprocessing. *Trends in Cognitive Sciences, 2*(12), 484-492.
21. Sutherland, G., & McNaughton, B. (2000). Memory trace reactivation in hippocampal and neocortical neuronal ensembles. *Current Opinions in Neurobiology, 10*, 180-186.
22. Traub R., & Miles R (1991). *Neuronal networks of the hippocampus*. Cambridge, UK: Cambridge Univ. Press.
23. Treves A., Rolls E. (1994) Computational analysis of the role of the hippocampus in memory. *Hippocampus, 4*, 374-391.
24. Vertes, Robert P. and Eastman, K. E. (2000). The Case Against Memory Consolidation in REM sleep. *Behavioral and Brain Sciences, 23* (6).