Hebbian Learning and Temporary Storage in the Convergence-Zone Model of Episodic Memory

Michael Howe

Department of Computer Sciences, The University of Texas at Austin, Austin, TX 78712

mhowe@cs.utexas.edu

Risto Miikkulainen

Department of Computer Sciences, The University of Texas at Austin, Austin, TX 78712

risto@cs.utexas.edu

Abstract

The Convergence-Zone model shows how sparse, random memory patterns can lead to one-shot storage and high capacity in the hippocampal component of the episodic memory system. This paper presents a biologically more realistic version of the model, with continuously-weighted connections and storage through Hebbian learning and normalization. In contrast to the gradual weight adaptation in many neural network models, episodic memory turns out to require high learning rates. Normalization allows earlier patterns to be overwritten, introducing time-dependent forgetting similar to the hippocampus.

Key words: Convergence-zone; Episodic memory; Neural network;

1 Introduction

Several recent results suggest that the episodic memory system consists of two components: the hippocampal formation serves as a fast, temporary storage where the traces are created immediately as the experiences come in, and the neocortex organizes and stores the experiences for the lifetime of the individual (1; 3; 4; 7). The hippocampal component of this system has been studied in detail for a long time, and although much is known about its anatomy and

To appear in the *Proceedings of the 8th Annual Computational Neuroscience Meeting*, 1999.

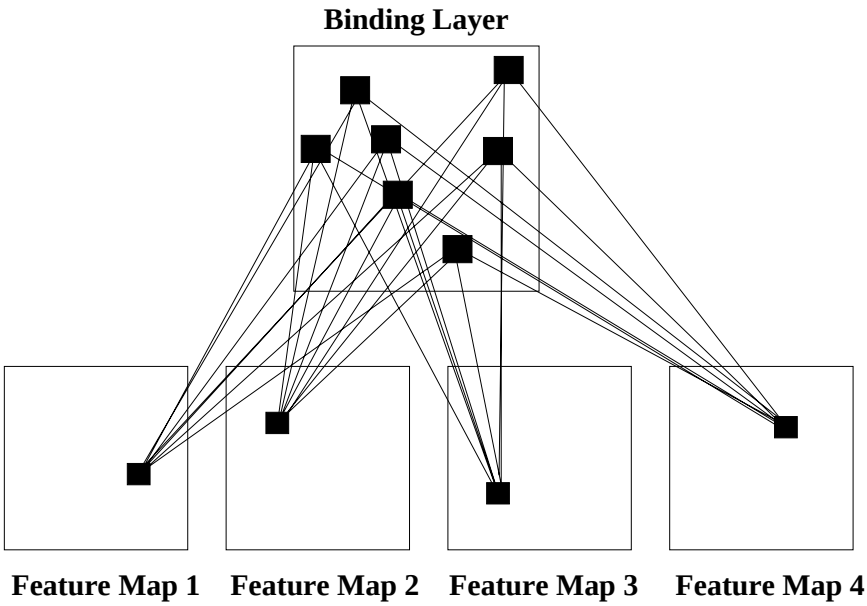

Fig. 1. **The Convergence-Zone model of hippocampal episodic memory.** The perceptual input pattern is stored in the bidirectional weights between the feature map units and the binding units. During retrieval, a partial input pattern will reactivate the binding pattern, which in turn reactivates the complete input pattern.

function, how exactly it manages to accurately store a large number of memory traces for several days is still not well understood.

Computational modeling can serve as an important tool in formulating and testing hypotheses about the hippocampal memory system. For example, the Convergence-Zone model (6, Figure 1) shows why the memory encoding areas can be much smaller than the perceptual maps, why they could consist of rather coarse computational units, and be only sparsely connected to the perceptual maps. In this model, an input to be stored is presented as a pattern of activation across a set of feature maps, representing high-level abstractions of sensory information. The memory is stored as a sparse, random pattern of activation in a binding layer, which is a convergence zone (2) representing the hippocampus. At each presentation, the connections between the active units in the feature maps and in the binding layer are turned on. When an incomplete pattern is presented on the feature maps, the binding layer pattern is activated, which in turn activates the complete pattern in the feature maps.

The Convergence-Zone model can answer several questions about the properties of sparse, random representations in the hippocampus, but it is an abstract, high-level model. The connections between the feature map and the binding layer of the convergence zone memory are represented as binary values, and learning occurs by switching a given set of connections from inactive to active. This paper focuses on a biologically more realistic implementation

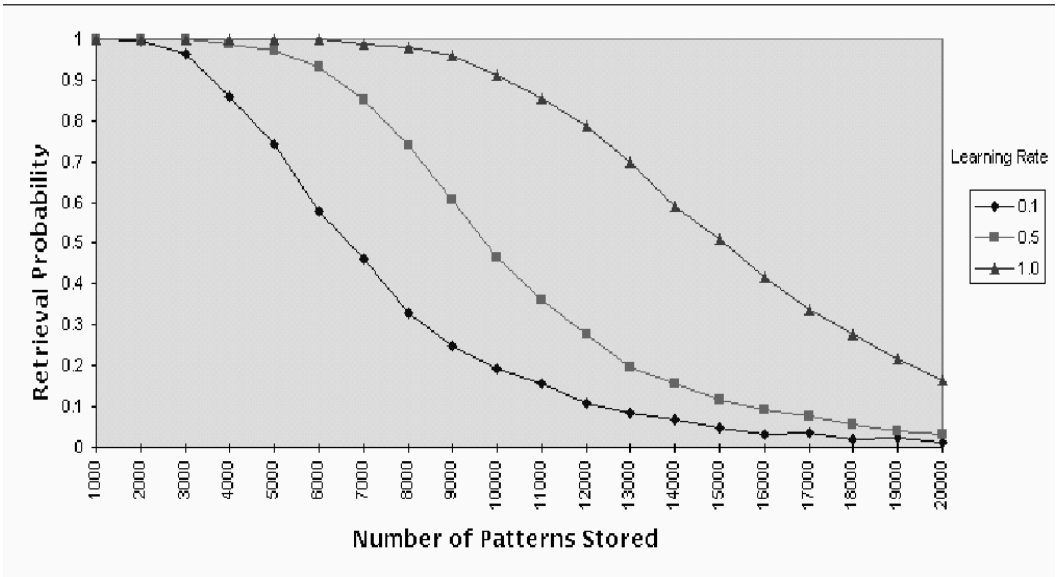

Fig. 2. **Capacity with different learning rates.** The probability of correct retrieval is shown as a function of total number of patterns stored. The plots are averages of 8 runs on a model consisting of 4 feature maps, each with 4000 units, and a binding layer of 200 units with binding patterns of 20 units. Higher learning rates result in better capacity.

of the model, with continuously-weighted connections and Hebbian weight adjustments. The model shows that episodic memory requires high learning rates, and that normalization can account for the temporary nature of the encoding of memories in the hippocampus.

2 Hebbian learning and one-shot learning

In the first experiment, at each presentation the connections between the active feature map units and the binding layer units were increased according to a given learning rate. The weights were initially 0, and were limited to [0, 1]. Whereas in the old version of the model the weights were turned from 0 to 1 in one shot, the idea was to check whether maximal retrieval accuracy, and thus largest capacity, could be achieved by using an intermediate learning rate between 0 and 1, and therefore intermediate weight values. Such a learning rate would allow fine tuning of the connection weights, and the model could then make subtle distinctions between various patterns stored in memory. To examine this factor we varied the learning rate for a given architecture (Figure 2).

All learning rates showed the same general trend: there was initially a high level of retrieval accuracy followed by a sharp decline as more patterns were stored. However, interestingly the drop off point was higher for higher rates,

i.e. the learning rates near 1.0 resulted in the highest capacity. Apparently the one-shot nature of episodic memory requires different kind of learning than is usually required of neural network models. It is not necessary to represent similarities between inputs; instead, it is necessary to store the patterns exactly as they are, and it is best done with maximal changes.

3 Normalization and forgetting

To prevent the weights from increasing without bounds, Hebbian learning is usually combined with normalization (instead of posing hard limits on the weight values). Normalization is biologically realistic due to limited resources of the neuron and decay of unused connections (5; 8), but it also results in an important effect in the episodic memory model: it introduces time-dependent forgetting. Without normalization, traces are sometimes lost because their representations overlap, but each trace is equally likely to be preserved or lost: it does not matter when they were stored. With normalization, the earlier traces gradually become more difficult to retrieve. This way the memory retains each trace for only a limited amount of time, as is the case in the hippocampus.

There are several ways of implementing normalization:

(1) The sum of output weights of feature map neurons can be kept constant
(2) All output weights of the active feature map units can be decayed by a constant factor
(3) The sum of input weights of the binding units can be kept constant
(4) The input weights of the active binding units can be decayed
(5) All weights in the system can be decayed at each time step. All these methods were implemented and tested, however, they resulted in very similar behavior.

Figure 3 shows the forgetting profiles for the first type of normalization. Time is in the x-axis, and the number of patterns stored is shown in the y-axis. The z-value shows how likely it is that the pattern stored at a particular time in the past will be retrieved after a given number of patterns have been stored. As can be seen from the graph, there is a limited time window during which the old patterns can be recalled, and a gradual drop off after that, much like in the hippocampus. Smaller learning rates generally make the drop off more gradual.

In the model so far, the weights have been bidirectional, representing the strength of association between a feature map unit and a binding unit. With normalization, it is possible that the weights in the two directions become different. Such a unidirectional model was also tested; the forward propagating

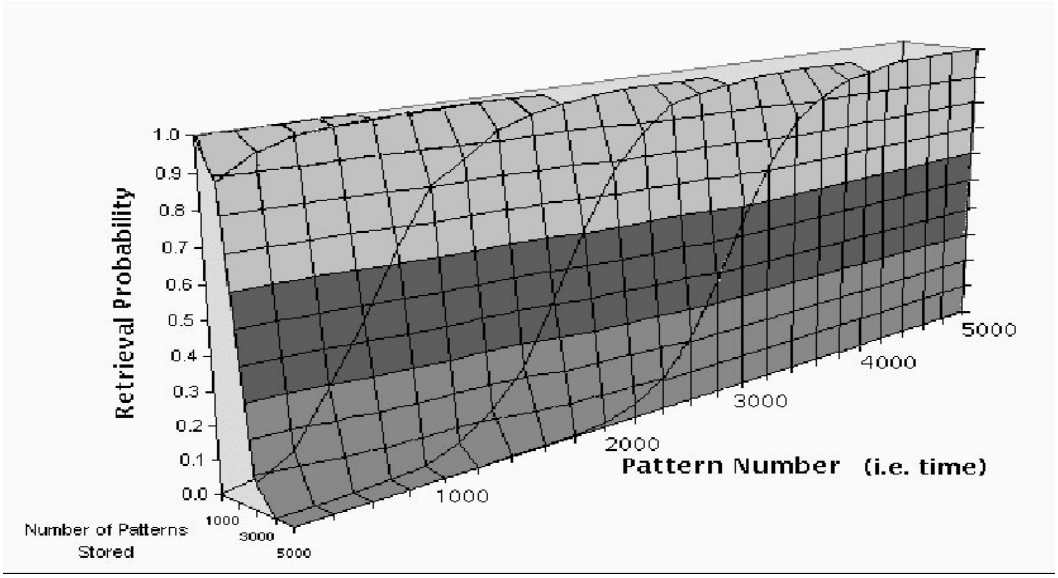

Fig. 3. **Forgetting based on normalization.** The same model was used as in Figure 2, with a learning rate of 1.0. Normalization was performed on the binding units, as in (3) above. The probability of correct retrieval is plotted as a function of the total number of patterns stored, and the time of storage. There is a temporal window when a pattern can be retrieved, and a sharp decline after that.

weights were normalized per active feature map unit, and the reverse weights were normalized per active binding layer unit. The results were quite similar to the bidirectional model, suggesting that bidirectionality is a valid approximation.

4 Conclusion

A biologically more realistic version of the convergence zone model of hippocampal episodic memory was presented. The effects of continually-weighted connections and Hebbian weight adaptation was tested, to better model the gradual increases and decreases of strengths possible with neural connections. The simulations show that the one-shot nature of episodic memory requires large learning rates, in contrast to standard neural network models. In addition, weight normalization allows earlier patterns to be overwritten, establishing a mechanism for temporary storage, similar to the function of the hippocampus.

References

[1] P. Alvarez and L.R. Squire, Memory consolidation and the medial temporal lobe: A simple network model, *Proceedings of the National Academy of Sciences of the USA*, **91** (1994) 7041-7045.

[2] A.R. Damasio, The brain binds entities and events by multiregional activation from convergence zones, *Neural Computation*, **1** (1989) 123-132.

[3] E. Halgren, Human hippocampal and amygdala recording and stimulation: Evidence for a neural model of recent memory, in: L Squire and N Butters, eds., *The Neuropsychology of Memory* (Guilford, New York, 1984) 165-182.

[4] J.L. McClelland, B.L. McNaughton, and R.C. O'Reilly, 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** (1995) 419-457.

[5] K.D. Miller and D.J.C. MacKay, The role of constraints in Hebbian learning, *Neural Computation*, **6** (1994) 100-126.

[6] M.Moll and R.Miikkulainen, Convergence-zone episodic memory: Analysis and simulations, *Neural Networks*, **10** (1997) 1017-1036.

[7] L.R. Squire, Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans, *Psychological Review*, **99** (1992) 195-231.

[8] G. G. Turrigiano, K. R Leslie, N.S. Desai, L.C. Rutherford, and S.B. Nelson, Activity-dependent scaling of quantal amplitude in neocortical neurons, *Nature*, **391** (1998) 845-846.