The Capacity of Convergence-Zone Episodic Memory

Mark Moll
Department of Computer Science
University of Twente
P.O. Box 217, 7500 AE Enschede
The Netherlands
moll@cs.utwente.nl

Risto Miikkulainen
Department of Computer Sciences
The University of Texas at Austin
Austin, TX 78712 USA
risto@cs.utexas.edu

Jonathan Abbey
Applied Research Laboratories
P.O. Box 8029
Austin, TX 78713 USA
broccol@arlut.utexas.edu

Abstract

Human episodic memory provides a seemingly unlimited storage for everyday experiences, and a retrieval system that allows us to access the experiences with partial activation of their components. This paper presents a neural network model of episodic memory inspired by Damasio's idea of Convergence Zones. The model consists of a layer of perceptual feature maps and a binding layer. A perceptual feature pattern is coarse coded in the binding layer, and stored on the weights between layers. A partial activation of the stored features activates the binding pattern which in turn reactivates the entire stored pattern. A worst-case analysis shows that with realistic-size layers, the memory capacity of the model is several times larger than the number of units in the model, and could account for the large capacity of human episodic memory.

Introduction

Human memory system can be divided into semantic memory of facts, rules, and general knowledge, and episodic memory that records the individual's day-to-day experiences Tulving (1972, 1983). Episodic memory is characterized by an extremely high capacity. New memories are formed every few seconds, and many of those persist in the memory for years, even decades (Squire 1987). Another significant characteristic of human memory is content-addressability. Most of the memories can be retrieved simply by activating a partial representation of the experience, such as a sound, a smell, or a visual image.

Although several artificial neural network models of episodic memory have been proposed (Hopfield 1982; Kanerva 1988; Kortge 1990; Miikkulainen 1992), they fall short of explaining the simultaneous huge capacity and content-addressability of human memory. For example in the Hopfield model of N units, $N/4 \log N$ patterns can be stored with a 99% probability of correct retrieval when N is large (Hertz, Krogh, & Palmer 1991; Keeler 1988; McEliece *et al.* 1986). This means that storing and retrieving, for example, 10^8 memories would require in the order of 10^{10} nodes and 10^{20} connections. Given that the human brain is estimated to have about 10^{11} neurons and 10^{15} synapses (Jessell 1991), this is clearly unrealistic.

Despite vast amount of research in human memory, no clear understanding has yet emerged on exactly where and how the memory traces are represented in the brain. There is evidence for both localized encoding and for distributed encoding (Squire 1987). Damasio (1989b, 1989a) proposed a general framework, based on observations of typical patterns of injury-related memory deficits, that can potentially account for much of the data. The main idea is that the memory system is organized in a hierarchy of associational regions, or convergence zones, with each region serving as a basis for higher-level associations. The hierarchy is grounded in the sensory modality regions, and becomes more abstract and general as one moves from the sensory cortical regions to the forebrain. The low-level and intermediate regions contain object representations, and the high-level regions contain representations for complete episodes, in terms of the lower-level entities.

This paper presents a new episodic memory model loosely based on the convergence zone idea. The model consists of a number of perceptual maps and a binding layer (a convergence zone). An episodic experience appears as a pattern of local activations across the perceptual maps, and is encoded as a coarse-coded (Rosenfeld & Touretzky 1989; Touretzky & Hinton 1988) pattern in the binding layer. The connections between the maps and the binding layer store the encoding so that the complete perceptual pattern can later be regenerated from partial activation. The details of the low-level neural implementation are left open in this paper. The goal is to analyze the behavior of the model at the functional level, and derive general results about its capacity and physical size.

A worst-case analysis of the model shows that: (1) with realistic-size maps and binding layer, the capacity of the convergence-zone memory is extremely high, exceeding the number of units in the model by a factor of 5; and (2) the majority of the neural hardware is required in the perceptual processing; the binding

To appear in the *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 1994.

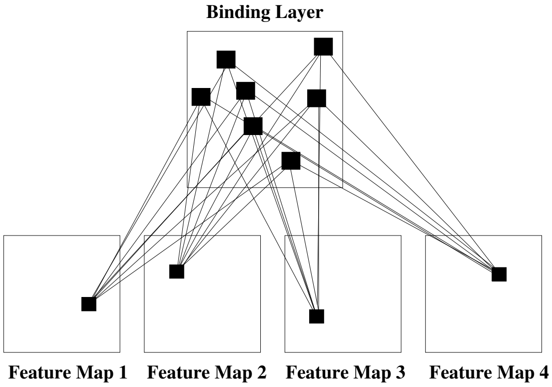


Figure 1: **Storage.** The weights on the connections between the appropriate feature units and the binding representation of the pattern are set to 1.

layer needs to be only a fraction of the size of the perceptual maps. Such results suggest how an extremely high capacity could be achieved in the human episodic memory with very little extra hardware beyond the perceptual maps.

Storage and Retrieval

The model consists of two layers of real-valued units (the feature map layer and the binding layer), and bidirectional binary connections between the layers (figure 1). Perceptual experiences are represented as vectors of feature values, such as color=`red`, shape=`round`, size=`small`. The values are encoded as units on the feature maps. There is a separate map for each feature domain, and each unit on the map represents a particular value for that feature. For instance, on the map for the color feature, the value `red` could be specified by turning on the unit in the lower-right quarter (figure 1). The feature map units are connected to the binding layer with bidirectional binary connections (i.e. the weight is either 0 or 1). An activation of units in the feature map layer causes a number of units to become active in the binding layer, and vice versa. In effect, the binding layer activation is a compressed, distributed encoding of the value-unit perceptual representation.

Initially, all connections are inactive at 0. A perceptual experience is stored in the memory through the feature map layer in three steps. First, those units that represent the appropriate feature values are activated at 1. Second, a subset of m binding units are randomly selected in the binding layer as the compressed encoding for the pattern, and activated at 1. Third, the weights of all the connections between the active units in the feature maps and the active units in the binding layer are set to 1 (figure 1). Note that only one presentation is necessary to store a pattern.

To retrieve a pattern, first all binding units are set to 0. The pattern to be retrieved is partially specified in the feature maps by activating a subset of its feature units. For example, in figure 2a the memory is cued with the two leftmost features. The activation propagates to the binding layer through all connections that have been turned on so far. The set of binding units that a particular feature unit turns on is called the binding constellation of that unit. All binding units in the binding encoding of the pattern to be retrieved are active at 2 because they belong to the binding constellation of both retrieval cue units. A number of other units are also activated at 1, because each cue unit takes part in representing multiple patterns, and therefore has several other active connections as well. Only those units active at 2 are retained; units with less activation are turned off (figure 2b).

The activation of the remaining binding units is then propagated back to the feature maps (figure 2c). A number of units are activated at various levels in each feature map, depending on how well their binding constellation matches the current pattern in the binding layer. Chances are that the unit that belongs to the same pattern than the cues has the largest overlap and becomes most highly activated. Only the most active unit in each feature map is retained, and as a result, a complete, unambiguous perceptual pattern is retrieved from the system (figure 2d).

Retrieval Errors

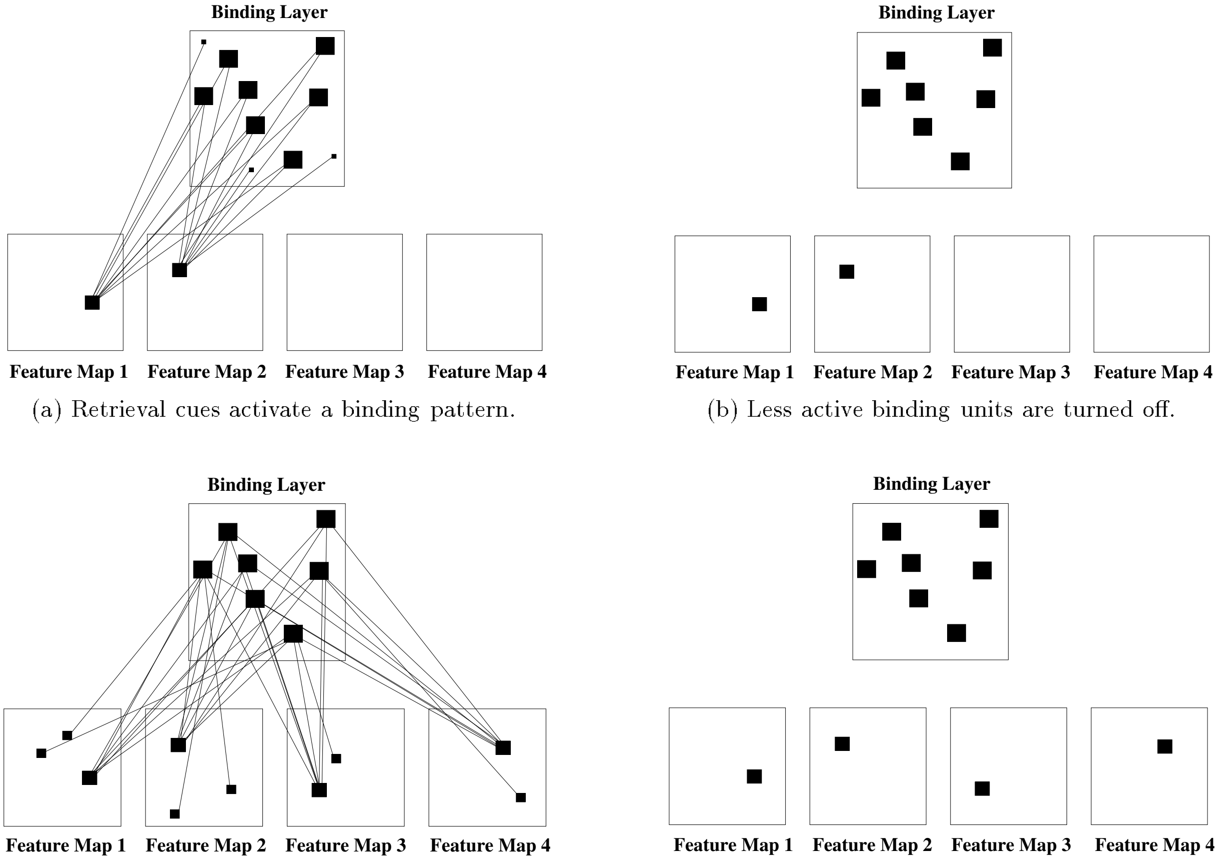
If there are n units in the binding layer and m units are chosen as a representation for a pattern, the number of possible different binding representations is equal to $\binom{n}{m}$. If n is sufficiently large and m is relatively small compared to n, this number is extremely large, suggesting that the convergence-zone memory could have a very large capacity.

However, due to the probabilistic nature of the storage and retrieval processes, there is always a chance that the retrieval will fail. The binding constellations of the retrieval cue units may overlap significantly, and several spurious units may be turned on at the binding layer. When the activation is propagated back to the feature maps, some random unit in a feature map may have a binding constellation that matches the spurious units very well. The "rogue" unit may receive more activation than the correct unit, and a wrong feature value may be retrieved. As more patterns are stored, the binding constellations of feature units become larger, and erroneous retrieval becomes more likely.

To determine the capacity of the convergence-zone memory, the chance of retrieval error must be computed. Below, a probabilistic formulation of the model is first given, and bounds for retrieval error are then computed.

Probabilistic Formulation

Let Z_i be the size of the binding constellation of a feature unit after i patterns have been stored on it and



(a) Retrieval cues activate a binding pattern.

(b) Less active binding units are turned off.

(c) Binding pattern activates feature units.

(d) Less active feature units are turned off.

Figure 2: **Retrieval.** A stored pattern is retrieved by presenting a partial representation as a cue. The size of the square indicates activation level of the unit.

let Y_i be its increase after storing the ith pattern on it. Obviously, $Y_1 = m$. To obtain the distribution of Y_i when $i > 1$, note that the new active connections belong to the intersection of a randomly chosen subset of m connections among all n connections of the unit, and its all remaining inactive connections (a set with $n - z$ elements, where z is the binding constellation at the previous step). Therefore, $Y_i, i > 1$ is hypergeometrically distributed with parameters $m, n - z$, and n:

$$\mathrm{P}(Y_i = y | Z_{i-1} = z) = \binom{n-z}{y}\binom{z}{m-y} \Big/ \binom{n}{m}. \tag{1}$$

The constellation size Z_i is then given by

$$Z_i = \sum_{k=1}^{i} Y_k. \tag{2}$$

Let I be the number of patterns stored on a particular feature unit after p patterns have been stored in the entire memory. I is binomially distributed with parameters p and $\frac{1}{f}$, where f is the number of units in a feature map:

$$I \sim \mathrm{B}(p, \frac{1}{f}). \tag{3}$$

Let Z be the binding constellation of a particular feature unit after p patterns have been stored in the memory. It can be shown that $\mathrm{E}(Z) = n(1-(1-\frac{m}{nf})^p)$. The binding constellation of a feature unit, given that at least one pattern has been stored on it, is denoted by Z'; obviously $\mathrm{E}(Z') > E(Z)$. The variable Z' can be used to denote the binding constellation of a retrieval cue, which necessarily must have been used once, assuming that the retrieval cues are valid. Let Z'_j be the binding constellation of the jth retrieval cue and let X_j be the number of units in the intersection of the first j retrieval cues. Then $X_1 = Z'_1$. To get X_j for $j > 1$, we remove from consideration the m units all retrieval cues necessarily have in common (because they belong to the same stored pattern), and randomly select $z - m$ units from the total set of $n - m$ units and see how many of them belong to the current intersection of $x_{j-1} - m$ units. This is a hypergeometric distribution with parameters $z-m, x_{j-1}-m$, and $n-m$:

$$\mathrm{P}(X_j = x_j | Z'_j = z, X_{j-1} = x_{j-1}) = \binom{x_{j-1}-m}{x_j-m}\binom{n-x_{j-1}}{z-x_j} \Big/ \binom{n-m}{z-m}. \quad (4)$$

The intersection is taken over the binding constellations of all j retrieval cues.

The number of units in common between a potential rogue unit and the j retrieval cues is denoted by R_{j+1} and is also hypergeometrically distributed, however with parameters z, x, and n because we cannot assume that the rogue unit has at least m units in common with the cues:

$$\mathrm{P}(R_{j+1} = r | Z = z, X_j = x) = \binom{x}{r}\binom{n-x}{z-r} \Big/ \binom{n}{z}. \quad (5)$$

The correct unit in a feature map where a retrieval cue was not presented will receive an activation X_{j+1}. The correct unit will be retrieved if $X_{j+1} > R_{j+1}$, which is usually the case because $\mathrm{E}(X_{j+1}) > \mathrm{E}(R_{j+1})$. In each feature map there are $(f-1)$ potential rogue units, so the conditional probability of successful retrieval is $(1-\mathrm{P}(R_{j+1} > X_{j+1} | X_{j+1}, Z, X_j))^{(f-1)}$, not addressing tie-breaking. Unfortunately, it is very difficult to compute p_{success}, the unconditional probability of successful retrieval, because the distribution functions of Z, X_j, X_{j+1} and R_{j+1} are not known. But it is possible to derive bounds for p_{success} and show that with reasonable values for n, m, f, and p, the memory is reliable.

Lower bound for memory capacity

Memory capacity can be defined as the maximum number of patterns that can be stored in the memory so that the probability of correct retrieval with a given number of retrieval cues is greater than α (a constant close to 1). In this section, worst-case bounds for the chance of successful retrieval will be derived. The analysis consists of three steps: (1) bounds for the number of patterns stored on a feature unit; (2) bounds for the binding constellation size; and (3) bounds for the intersections of binding constellations. Given particular values for the system parameters, it is then possible to give a lower bound for the capacity of the model.

1. Number of patterns stored on a feature unit. Since I has a binomial distribution (with parameters p and $\frac{1}{f}$), Chernoff bounds can be applied:

$$\mathrm{P}(I \le (1-\delta)\frac{p}{f}) \le \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\frac{p}{f}}, 0 < \delta < 1, \quad (6)$$

$$\mathrm{P}(I \ge (1+\delta)\frac{p}{f}) \le \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\frac{p}{f}}, \qquad \delta > 0. \quad (7)$$

The formal parameter δ determines the tradeoff between the tightness of the bounds and the probability of satisfying them.

2. Size of the binding constellation. Instead of choosing exactly m different units for the binding representation of each pattern, let us select k not-necessarily-distinct units in such a way that the *expected* number of different units is m. This will make the analysis easier at the cost of larger variance, so that the bounds derived will also be valid for the actual process.

Let us assume i patterns are stored on a unit, which is equivalent of selecting ki units from the binding constellation at random. Let Z_v be the expected size of the binding constellation after v units have been selected. Then

$$Z_v = \tilde{Z} + (n-\tilde{Z})(1-(1-\frac{1}{n})^{ki-v}), \quad (8)$$

where $\tilde{Z}$ is the size of the binding constellation formed by the first v selected units. Now, $\mathrm{E}(Z_v|Z_{v-1}) = Z_{v-1}$, and the sequence of variables $Z_0, \ldots, Z_{ki}$ is a martingale. Moreover, it can be shown that $|Z_v - Z_{v-1}| \le 1$, and bounds for Z can be obtained from Azuma's inequality (see e.g. Alon & Spencer 1992):

$$\mathrm{P}(Z \le n(1-(1-\frac{1}{n})^{ki_l}) - \lambda\sqrt{ki_l}) \le e^{-\lambda^2}, \quad (9)$$

$$\mathrm{P}(Z \ge n(1-(1-\frac{1}{n})^{ki_u}) + \lambda\sqrt{ki_u}) \le e^{-\lambda^2}, \quad (10)$$

where i_l is the lower bound for I obtained from equation 6, and i_u the upper bound from equation 7. Similar bounds can be derived for Z'.

3. Intersection of binding constellations. The process of forming the intersection of j binding constellations incrementally one cue at a time can also be formulated as a martingale process. Let X_j denote the expected number of elements in the intersection of two sets, after the first j elements of the first set have been checked (the elements of the second set are assumed to be known at all times). Then

$$X_j = \tilde{X} + \frac{(n_1-j)(n_2-\tilde{X})}{n-j}, \quad (11)$$

where $\tilde{X}$ is the number of elements in the intersection of the second set and the set formed by the first j elements of the first set, and n_1, n_2 and n are the sizes of the first, second, and the superset. If n_1 and n_2 are both smaller than $\frac{1}{2}n$, Azuma's inequality can be applied. Taking the intersection of the previous step as the first set, the binding constellation of the jth cue as the second set, and the binding layer as the common superset, this approach gives us the following upper bound for X_j:

$$\mathrm{P}(X_j \ge \frac{(x_{j-1,u}-m)(z'_u-m)}{(n-m)} + m + \lambda\sqrt{x_{j-1,u}-m}) \le e^{-\lambda^2/2}, \qquad \lambda > 0, \quad (12)$$

where z'_u and $x_{j-1,u}$ are upper bounds for Z' and X_{j-1} and are assumed to be less than $\frac{1}{2}n$. When X_j is at its upper bound, a potential rogue unit has the largest chance of taking over. In this case, R_{j+1} has the upper bound

$$\mathrm{P}(R_{j+1} \geq \frac{x_{j,u} z_u}{n} + \lambda\sqrt{x_{j,u}}) \leq e^{-\lambda^2/2}, \quad \lambda > 0, \quad (13)$$

where z_u and $x_{j,u}$ are upper bounds for Z and X_j. A lower bound for X_{j+1} while using an upperbound for X_j is then given by

$$\mathrm{P}(X_{j+1} \leq \frac{(x_{j,u} - m)(z_l - m)}{(n-m)} + m \\ -\lambda\sqrt{x_{j,u} - m}) \leq e^{-\lambda^2/2}, \qquad \lambda > 0. \qquad (14)$$

If the resulting lower bound is smaller than m, m can be used instead.

The above analysis ignores correlations between binding constellations. The correlations originate from storing the same partial pattern multiple times and tend to increase the size of the intersections. The chance that two random patterns have more than one feature in common in j features is equal to $(1-(1+\frac{j}{f-1})(1-\frac{1}{f})^j)$, which is negligible for sufficiently large values of f.

We can now use equations 6–14 to derive a lower bound for the probability of successful retrieval with given system parameters n, m, F, j, f, and p. The retrieval is successful if $r_{j+1,u}$, the upper bound for R_{j+1}, is lower than $x_{j+1,u}$, the lower bound for X_{j+1}. Under this constraint, the probability that none of the variables in the analysis exceeds its bounds is a lower bound for successful retrieval.

Obtaining the upper bound for X_j involves bounding $3j-1$ variables: I and Z' for the j cues and X_j for the $j-1$ intersections. Computing $x_{j+1,l}$ and $r_{j+1,u}$ each involve bounding 3 variables (I, Z, and X_{j+1}; I, Z', and R_{j+1}). There are $F-j$ maps, each with one $x_{j+1,l}$ bound and $f-1$ different $r_{j+1,u}$ bounds (one for each rogue unit). The total number of bounds is therefore $3j-1+3f(F-j)$. Setting the righthand sides of the inequalities 6–14 equal to a small constant β, a lower bound for successful retrieval is obtained:

$$p_{\text{success}} > 1 - (3j - 1 + 3f(F-j))\beta. \qquad (15)$$

For example, assuming each unit in the model corresponds to a vertical column in the cortex, it is reasonable to assume feature maps with 10^6 computational units (Sejnowski & Churchland 1989). We can further assume that the system has 15 feature maps, 10 of which is used to cue the memory, and the binding layer consists of 10^5 units, with 150 used for each binding pattern. Assuming full connectivity between the feature units and the binding units, there are 1.5×10^{12} connections in the system.

If we store 0.85×10^8 patterns in the memory, z'_u and $x_{j-1,u}$ are less than $\frac{1}{2}n$, the chance of partial overlap of more than 1 feature is less than 0.45×10^{-10}, and the analysis above is valid. Setting $\beta = 0.5 \times 10^{-9}$ yields bounds $r_{j+1,u} < x_{j+1,l}$ with $p_{\text{success}} > 99\%$. In other words, 0.85×10^8 memories can be stored in the memory with 99% probability of successful retrieval. Such a capacity is approximately equivalent of storing one new memory every 17 seconds for 70 years, 16 hours a day.

Conclusion

Mathematical analysis shows that an extremely high number of episodes can be stored in the convergence-zone memory with reliable content-addressable retrieval. Moreover, the convergence zone itself requires only a tiny fraction of the hardware required for perceptual representation. These results provide a possible explanation for why human memory appears almost unlimited, and why memory areas appear small compared to the areas devoted to low-level perceptual processing.

The model makes use of the combinatorics and the clean-up properties of coarse coding in a neurally-inspired architecture. The storage capacity of the model appears to be at least two orders of magnitude higher than that of the Hopfield model with the same number of units, while using two orders of magnitude fewer connections. However, direct comparison is difficult because the stored patterns in the Hopfield model are much larger (contain more information), and its $N/4 \log N$ capacity result only indicates how many patterns are stable instead of estimating the probability of correct retrieval with a partial pattern as a cue.

The convergence-zone episodic memory model could be extended to make it more accurate as a model of actual neural processes. For instance, lateral inhibitory connections between units within a feature map could be added to select the unit with the highest activity. A similar extension could be applied to the binding layer; instead of only one unit multiple units should stay active. A variation of the Hebbian learning mechanism (Hebb 1949; Miller & MacKay 1992) could be used to implement the storage mechanism. Such research could lead to a practical implementation of the convergence zone memory, and perhaps even to a hardware implementation. Another important research direction is to analyze the behavior of the model as a psychological model, that is, to observe and characterize its memory interference effects and compare them with experimental results on human episodic memory.

Acknowledgements

We would like to thank Greg Plaxton for pointing us to martingale analysis on this problem. This research was supported in part by NSF grant #IRI-9309273 to the second author.

References

Alon, N., and Spencer, J. H. 1992. *The Probabilistic Method*. New York: Wiley.

Damasio, A. R. 1989a. The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation* 1:123–132.

Damasio, A. R. 1989b. Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition* 33:25–62.

Hebb, D. O. 1949. *The Organization of Behavior: A Neuropsychological Theory*. New York: Wiley.

Hertz, J.; Krogh, A.; and Palmer, R. G. 1991. *Introduction to the Theory of Neural Computation*. Reading, MA: Addison-Wesley.

Hopfield, J. J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA* 79:2554–2558.

Jessell, E. R. K. 1991. Nerve cells and behavior. In Kandel, E. R.; Schwartz, J. H.; and Jessell, T. M., eds., *Principles of Neural Science*. Elsevier. 18–32.

Kanerva, P. 1988. *Sparse Distributed Memory*. Cambridge, MA: MIT Press.

Keeler, J. D. 1988. Comparison between Kanerva's SDM and Hopfield-type neural networks. *Cognitive Science* 12:299–329.

Kortge, C. A. 1990. Episodic memory in connectionist networks. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, 764–771. Hillsdale, NJ: Erlbaum.

McEliece, R. J.; Posner, E. C.; Rodemich, E. R.; and Venkatesh, S. S. 1986. The capacity of the hopfield associative memory. *IEEE Transactions on Information Theory* 33:461–482.

Miikkulainen, R. 1992. Trace feature map: A model of episodic associative memory. *Biological Cybernetics* 66:273–282.

Miller, K. D., and MacKay, D. J. C. 1992. The role of constrains in Hebbian learning. CNS Memo 19, Computation and Neural Systems Program, California Institute of Technology, Pasadena, CA.

Rosenfeld, R., and Touretzky, D. S. 1989. A survey of coarse-coded symbol memories. In Touretzky, D. S.; Hinton, G. E.; and Sejnowski, T. J., eds., *Proceedings of the 1988 Connectionist Models Summer School*, 256–264. San Mateo, CA: Morgan Kaufmann.

Sejnowski, T. J., and Churchland, P. S. 1989. Brain and cognition. In Posner, M. I., ed., *Foundations of Cognitive Science*. Cambridge, MA: MIT Press. chapter 8, 315–356.

Squire, L. R. 1987. *Memory and Brain*. Oxford, UK; New York: Oxford University Press.

Touretzky, D. S., and Hinton, G. E. 1988. A distributed connectionist production system. *Cognitive Science* 12:423–466.

Tulving, E. 1972. Episodic and semantic memory. In Tulving, E., and Donaldson, W., eds., *Organization of Memory*. New York: Academic Press. 381–403.

Tulving, E. 1983. *Elements of Episodic Memory*. Oxford, UK; New York: Oxford University Press.