# Using Semi-Distributed Representations to Overcome Catastrophic Forgetting in Connectionist Networks

Robert M. French
Center for Research on Concepts and Cognition
Indiana University
510 North Fess
Bloomington, Indiana 47408
e-mail: french@cogsci.indiana.edu

## Abstract

In connectionist networks, newly-learned information can completely destroy previously-learned information unless the network is continually retrained on the old information. This behavior, known as catastrophic forgetting, is unacceptable both for practical purposes and as a model of mind. This paper advances the claim that catastrophic forgetting is a direct consequence of the overlap of the system's distributed representations and can be reduced by reducing this overlap. A simple algorithm is presented that allows a standard feedforward backpropagation network to develop semi-distributed representations, thereby significantly reducing the problem of catastrophic forgetting.

## Introduction

Catastrophic forgetting is the inability of a neural network to retain old information in the presence of new. New information destroys old unless the old information is continually relearned by the net. McCloskey & Cohen [1990] and Ratcliff [1989] have demonstrated that this is a serious problem with connectionist networks. A related problem is that connectionist networks are not sensitive to overtraining. A network trained 1000 times to associate a pattern A with a pattern A' will forget that fact just as quickly as would a network trained on that association for 100 cycles. Clearly, this behavior is unacceptable as a model of mind, as well as from a purely practical standpoint. Once a network has thoroughly learned a set of patterns, it should be able to learn a completely new set and still be able to recall the first set with relative ease. In this paper I will suggest that catastrophic forgetting arises because of the overlap of distributed representations and I will present an algorithm that will allow a standard feedforward backpropagation (FFBP) network to overcome to a significant extent the problems of catastrophic forgetting and insensitivity to overtraining.

## Catastrophic forgetting and the overlap of representations

I suggest the following relation between catastrophic forgetting and representations in a distributed system:

> Catastrophic forgetting is a direct consequence of the overlap of distributed representations and can be reduced by reducing this overlap.

Very local representations will not exhibit catastrophic forgetting because there is little interaction among representations. Consider the extreme example of a look-up table where there is no overlap at all among representations. There is no catastrophic forgetting; new information can be added without interfering at all with old information. However, because of its completely local representations, a look-up table lacks the all-important ability to generalize.

At the other extreme are fully distributed networks where there is considerable interaction among representations. This interaction is responsible for the networks' generalization ability. On the other hand, these networks are severely affected by catastrophic forgetting.

The moral of the story is that you can't have it both ways. A system that develops highly distributed representations will be able to generalize but will suffer from castastrophic forgetting; conversely, a system that develops very local representations will not suffer from catastrophic forgetting, but will lose some of its ability to generalize. The challenge is to develop systems capable of producing semi-distributed representations that are local enough to overcome catastrophic forgetting yet that are sufficiently distributed to nonetheless allow generalization.

In what follows, I will examine two distributed systems that do not suffer from catastrophic forgetting. Both of these systems work because their representations are not fully distributed over the entire memory, but rather are semi-distributed and hence exhibit limited representation overlap, at least prior to memory saturation. Finally, I will present a simple method that allows standard layered

feedforward backpropagation networks to develop semi-distributed representations in the hidden layer. Not only does this method appear to dramatically reduce catastrophic forgetting but it also allows the system's representations to partially reflect the degree to which a particular pattern has been learned. Even after a particular pattern has been learned, overlearning continues to modify connection weights in such a way that unlearning of the pattern will be made more difficult.

## Two examples of semi-distributed representations

I will briefly examine two systems that produce semi-distributed representations. In both systems, assuming that they are not saturated, there is little overlap of the representations produced. For this reason, they exhibit little catastrophic forgetting.

## Sparse Distributed Memory

Sparse Distributed Memory (hereafter, SDM [Kanerva 1988]) is an auto-associative, content-addressable memory typically consisting of one million 1000-bit "memory locations". The memory is called "sparse" because it uses only one million locations out of a possible $2^{1000}$ (i.e., $10^6$ of approximately $10^{300}$ possible locations). At each of these locations there is a vector of 1000 integers, called "counters". New data are represented in the system as follows: If we wish to write a particular 1000-bit string to this memory, we select all memory locations that are within a Hamming distance of 450 bits of the write address. This gives us approximately 1000 locations (i.e. 0.1% of all of the entire address space). Wherever there is a 1 in the bit-string to be written to memory, we increment the corresponding counter in each of the vectors at the 1000 memory locations; wherever there is a 0, we decrement the corresponding counter. This is clearly a semi-distributed representation of the input data: storage of the bit-string is distributed over 1000 different memory locations but these 1000 memory locations account for a mere 0.1% of the total available memory.

This system can easily store new information without interfering with previously stored information as long as the representations do not overlap too much. As soon as the memory starts to become saturated (at somewhat less than 100,000 words written to memory), there is interference among representations, and learning new information begins to interfere with the old. In this case, not only is there forgetting of the old information but the new information cannot be stored either.

## ALCOVE

ALCOVE [Kruschke 1990] is a computer memory model based on Nosofsky's exemplar memory model [Nosofsky 1984]. This model does not suffer from the phenomenon of catastrophic forgetting noted by Ratcliff and McCloskey & Cohen. As we will see, ALCOVE, like SDM, uses semi-distributed representations.

ALCOVE is a three-layer feed-forward network in which the activation of a node in the hidden layer is inversely exponentially proportional to the distance between the hidden node position and the input stimulus position. The hidden layer can be regarded as a "covering" of the input layer. The inverse exponential activation function has the effect of producing a localized receptive field around each hidden node, causing it to respond only to a limited part of the input field. This kind of localization does not exist in standard FFBP networks. This system therefore represents its inputs in a semi-distributed manner, with only a few hidden nodes taking part in the representation of a given input.

The architecture of ALCOVE is such that the representation of new inputs, especially of new inputs that are not close to already-learned patterns, will not overlap significantly with the old representations. This means that the set of weights that produced the old representations will remain largely unaffected by new input.

As in SDM, the representations in ALCOVE are also somewhat distributed, conferring on the system its ability to generalize. When the width of the receptive fields at each node is increased, thereby making each representation more distributed and causing greater overlap among representations, the amount of interference among representations increases.

## Semi-distributed representations in FFBP networks

If catastrophic forgetting could be reduced, the order in which inputs are presented to the network would be less important. Training could be done either sequentially or concurrently. In other words, the artificial constraint of requiring training data to be presented to the network in an interleaved fashion could be relaxed. If, in addition, the representations also reflected the amount of training required to produce them, it might be possible to produce a system that would better model overlearning than standard FFBP networks. An initial attempt to reduce catastrophic forgetting with semi-distributed representations by differentially modifying the learning rates of the connections in the network was described in [French & Jones 1991]. While this technique gave promising results on very small networks, it failed to scale up to larger networks. The algorithm presented below, using a different technique, allows semi-distributed representations to evolve that significantly reduce catastrophic forgetting.

## Activation overlap and representational interference in FFBP networks

Catastrophic forgetting is closely related to the much-studied phenomenon of crosstalk. The discussion of crosstalk has traditionally involved the capacity of a network to store information [Willshaw

1981]: above a certain capacity, distributed networks can no longer store new information without destroying old. In standard backpropagation models, there is a much more serious problem. As things currently stand, FFBP networks will not work at all without artificially interleaved training sets. Even when the network is nowhere near its theoretical storage capacity, learning *a single new input* can completely disrupt all of the previously learned information. Catastrophic forgetting is crosstalk with a vengeance.

A feedforward backpropagation network represents its inputs as activation patterns of units in the hidden layer. The amount of interaction among representations will be measured by their degree of "activation overlap". The activation overlap of a number of representations in the hidden layer is defined as their average shared activation over all of the units in the hidden layer. For example, if there are four hidden units and the representation for one input is (0.2, 0.1, 0.9, 0.1) and for a second is (0.2, 0.0, 1.0, 0.2), we calculate activation overlap by summing the smaller of the two activations (the "shared" activation) of each unit and averaging over all of the units. Here the activation overlap would be $(0.2 + 0.0 + 0.9 + 0.1)/4 = 0.3$.

I suggest that the amount that two representations interfere with one another is directly proportional to their amount of activation overlap. For example, consider the two following activation patterns: (1, 0, 0, 0) and (0, 0, 1, 0). Their activation overlap is 0. Regardless of the weights of the connections between the hidden layer and the output layer, there will be no interference in the production of two separate output patterns. But as activation overlap increases, so does the level of interference.

Therefore, if we can find a way to coax the network to produce representations with as little activation overlap as possible, we should be able to significantly reduce catastrophic forgetting.

## Sharpening the activation of hidden units

A technique that I call "activation sharpening" will allow an FFBP system to gradually develop semi-distributed representations in the hidden layer. Activation sharpening consists of increasing the activation of some number of the most active hidden units by a small amount, slightly decreasing the activation of the other units in a similar fashion, and then changing the input-to-hidden layer weights to accommodate these changes. The new activation for nodes in the hidden layer is calculated as follows:

$A_{new} = A_{old} + \alpha(1 - A_{old})$ for the nodes to be sharpened;

$A_{new} = A_{old} - \alpha A_{old}$ for the other nodes;

where $\alpha$ is the sharpening factor.

The idea behind this is the following. Nodes whose activation values are close to 1 will have a far more significant effect on the output, on average, than nodes with activations close to 0. If the system could evolve representations with a few highly activated nodes, rather than many nodes with average activation levels, this would reduce the average amount of activation overlap among representations. This should result in a decrease in catastrophic forgetting. In addition, because sharpening occurs gradually over the course of learning and continues even after a particular association has been learned, the representations developed will reflect the amount of training that it took to produce them.

Let us consider one-node sharpening. On each pass we find the most active node, increase its activation slightly and decrease the activations of the other nodes. To preserve these changes we then backpropagate the difference between the pre-sharpened activation and the sharpened activation to the weights between the input layer and the hidden layer. Here are the details of this activation sharpening algorithm for *k*-node sharpening:

- Perform a forward-activation pass from the input layer to the hidden layer. Record the activations in the hidden layer;
- "Sharpen" the activations of *k* nodes;
- Using the difference between the old activation and the sharpened activation on each node as "error", backpropagate this error to the input layer, modifying the weights between the input layer and the hidden layer appropriately;
- Do a full forward pass from the input layer to the output layer.
- Backpropagate as usual from the output layer to the input layer;
- Repeat.

## Alternative Implementations of the Algorithm

The experiments described below were run using the "backpropagation-and-a-half" algorithm described above. However, for one-node sharpening, lateral inhibitory links among the nodes of the hidden layer might achieve a similar effect. These inhibitory links would allow the node with the highest activation to damp the activation of the other less active nodes, thereby effectively "sharpening" the activation of the most active node with respect to the others. It is less clear how sharpening of two or more nodes could be achieved in a straightforward manner by means of lateral inhibitory connections.

## Experimental results

The experiments consisted of training (and overtraining) an 8–8–8 feedforward backpropagation network on a set of eleven associations. The learning rate was 0.2 and momentum 0.9. The network was then presented with a new association. After this new association had been learned, one of the associations from the first set was chosen and tested to see how well the system remembered it. On the first presentation of this previously learned association,

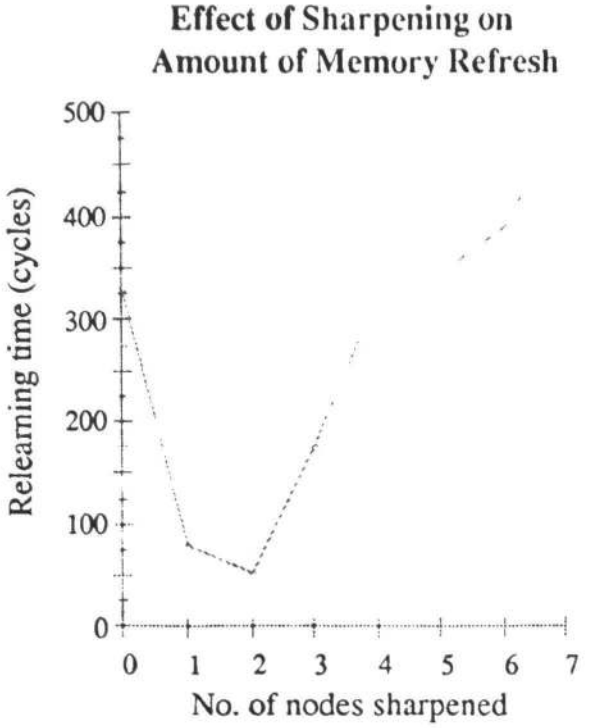


*Figure 1a*

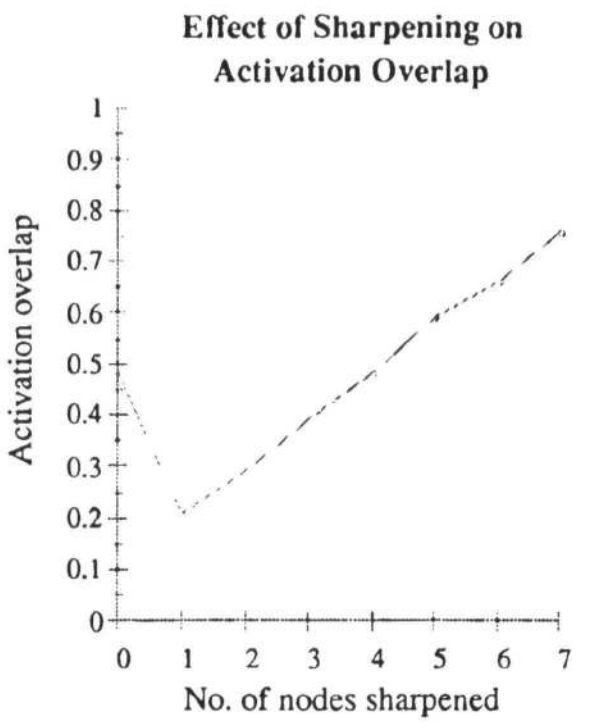


*Figure 1b*

the network invariably did very badly. The maximum error over all output nodes was almost always greater than 0.95 and the average error greater than 0.5. The amount of memory refresh required for a standard backpropagation network to relearn this association was recorded and compared to a network with one-node, two-node, three-node, etc. sharpening. In each case the sharpening factor was 0.2. The results are given in Figure 1a. (Note: 0-node sharpening is standard backpropagation.) It can be seen that one-node, two-node and three-node sharpening perform dramatically better than a standard FFBP network.

Over twenty separate runs, the standard FFBP network required an average of 330 cycles to relearn the previously-learned association. This figure dropped to 81 cycles for one-node sharpening and to 53 cycles for two-node sharpening. (Note: all runs were terminated at 500 cycles.) When the activations of three or more nodes were sharpened, the amount of relearning began to rise again. With three-node sharpening 175 cycles were required. With four-node (326 cycles) and five-node (346 cycles) sharpening, the modified system does no better than standard backpropagation. Above this, it does significantly worse. [Figure 1a]

The two graphs in Figures 1a and 1b suggest that amount of memory refresh required varies directly with the amount of activation overlap among representations. Figure 1b shows the amount of activation overlap of the eleven originally-learned inputs with various degrees of activation-sharpening. (As before "0 nodes sharpened" indicates standard backpropagation.) In general, the less activation overlap, the less the catastrophic forgetting as measured by the number of cycles required to relearn a previously-learned pattern.

In Figure 2 we can see the effect of this sharpening algorithm on the representations of one association. For each of twenty runs, the activation patterns on the hidden nodes at the end of the initial training period were recorded. The nodes in each of the twenty runs were sorted according to their activation levels and these figures were then averaged. As might be expected, for standard backpropagation the distribution of activations over the eight nodes was approximately uniform. This gives an activation profile from the most active nodes to least active nodes of approximately constant slope. However, the result of one-node sharpening is quite dramatic; one of the eight nodes was much more active than the other seven. The same phenomenon can be observed for the other experiments where two or more nodes where sharpened.

## Why does activation sharpening work?

Let us examine why activation sharpening reduces catastrophic forgetting. Consider two-node sharpening. As the system learns the first set of associations, it develops a set of sharpened representations in the hidden layer. A new association is then presented to the network. Activation sharpening immediately starts to coax the new representation into a sharpened form where two of the eight hidden nodes are highly active and six are not. Thus, very early on, the newly developing representation will have less chance of

## Effect of Sharpening on Hidden-Layer Activation Profiles

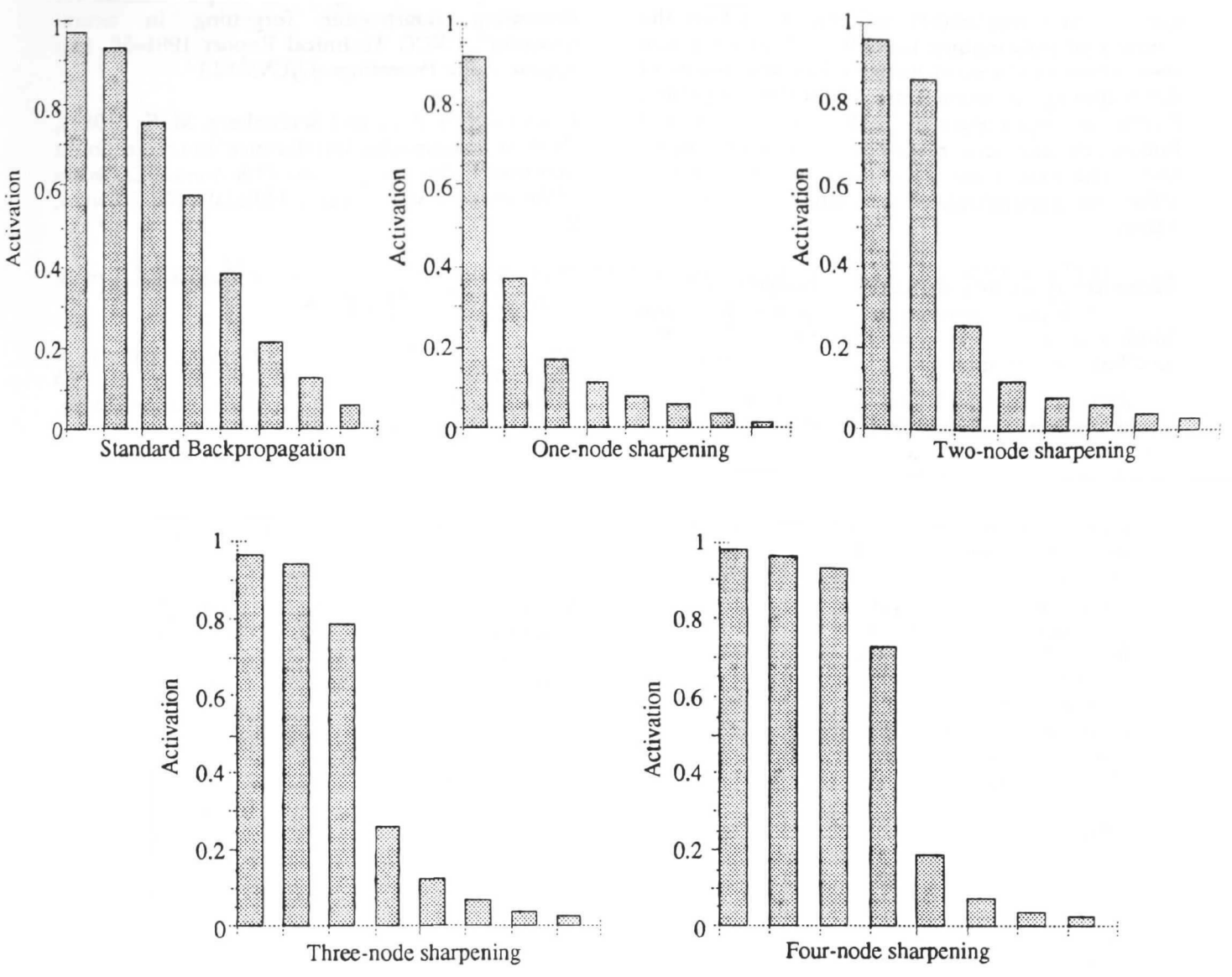


*Figure 2*

activation overlap with the already formed representations than in standard backpropagation, where the activation is spread out over all eight nodes.

Sharpened activation patterns interfere less with the weights of the network than unsharpened ones. The reason for this has to do with the way the backpropagation algorithm changes weights. When the activation of a node is near zero, the weight changes of the links associated with it are small. Thus, if a significant number of the nodes in the new representation have a very low activation, then the weights on the connections to and from that node will be modified much less, on average, than the weights associated with a highly active node. Therefore, the only representations significantly affected by the new representation will be those in which highly active nodes overlap. Consequently, if we reduce the probability of this overlap by activation sharpening, there will be a decrease in the amount of disruption of the old weights and catastrophic interference will be reduced.

The idea is to sharpen new activation patterns as quickly as possible, thereby decreasing their potential to interfere with already learned patterns. Keeping the learning rate low ($\leq 0.2$) with a relatively high sharpening factor (0.2) allows new activation patterns to become sharpened before they have a chance to do much damage to previously-learned weights. Preliminary experiments in fact indicate that as the learning rate is decreased with the sharpening factor held constant, catastrophic forgetting decreases.

It seems likely that semi-distributed

representations will cost the network some of its ability to generalize. Optimal generalization depends on as much information as possible taking part in mapping from the input space to the output space. Any mechanism tending to reduce the amount of information brought to bear on a new association would most likely reduce the quality of the mapping. In some sense, activation sharpening forces the input data through a representational bottleneck and this results in information being lost. The extent and severity of this loss and its effect on generalization is a subject of ongoing study.

## How many nodes should be sharpened?

This is an open question. For $n$ nodes in the hidden layer, the answer might be $k$ where $k$ is the smallest integer such that ${}_nC_k$ is greater than the number of inputs. In other words, a sufficient number of nodes should be sharpened to allow the existence of enough distinct sharpened representations to cover the input space. To minimize the activation overlap, the least such sufficient number of sharpened nodes should be chosen. If the number of input patterns to be learned is not known in advance, it might be reasonable to sharpen approximately log $n$ nodes. This estimate is based on work on crosstalk [Willshaw 1981]. This work indicates that in a distributed memory crosstalk can be avoided when the number of active units for each input pattern is proportional to the logarithm of the total number of units. It would seem reasonable to apply this result to the sharpening of hidden-unit activations.

## Conclusion

In this paper I have argued that catastrophic forgetting in distributed systems is a direct consequence of the amount of overlap of representations in that system. I have further suggested that the trade-off between catastrophic forgetting and generalization is inevitable. It is claimed that one way to maintain generalization capabilities while reducing catastrophic forgetting is to use semi-distributed representations. To this end, I presented a simple method to allow a feedforward backpropagation network to dynamically evolve its own semi-distributed representations.

## Acknowledgments

I would like to thank Mark Weaver for his invaluable assistance with the ideas and emphasis of this paper. I would also like to thank David Chalmers, Terry Jones, and the members of CRCC and SESAME for their many helpful comments.

## Bibliography

Feldman, J. A., [1988], "Connectionist Representation of Concepts", In *Connectionist Models and Their Implications*, Waltz, D. and Feldman, J. (eds.), 341–363.

French, R. M. and Jones, T. C., [1991], "Differential hardening of link weights: A simple method for decreasing catastrophic forgetting in neural networks", CRCC Technical Report 1991–50. [To appear in the *Proceedings of IJCNN-91.*]

Hetherington, P. A. and Seidenberg, M. S., [1989], "Is there 'catastrophic interference' in connectionist networks?", *Proceedings of the 11th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum, 26–33.

Kanerva, Pentti, [1988], *Sparse Distributed Memory*, Cambridge, MA: MIT Press.

Kortge, Chris A., [1990]. "Episodic Memory in Connectionist Networks", *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum, 764-771.

Kruschke, J. K., [1990], "ALCOVE: A exemplar-based connectionist model of category learning", Indiana University Cognitive Science Research Report 19, February 22, 1991.

McCloskey, M. and Cohen, N. J., [1989], "Catastrophic interference in connectionist networks: The sequential learning problem", *The Psychology of Learning and Motivation*, Vol. 24, 109-165.

Nosofsky, R. M., [1984], "Choice, similarity and the context theory of classification", *J. Exp. Psych. Learning, Memory and Cognition*, Vol. 10, 104-114.

Ratcliff, R., [1990], "Connections models of recognition memory: Constraints imposed by learning and forgetting function", *Psychological Review*, Vol. 97, 285-308.

Sloman, S. and Rumelhart, D., [1991], "Reducing interference in distributed memories through episodic gating", In A. Healy, S. Kosslyn, and R. Shiffrin (eds.), *Essays in Honor of W. K. Estes* (in press).

Weaver, M., [1990], "An active symbol connectionist model of concept learning" (unpublished manuscript).

Willshaw, D., [1981], "Holography, associative memory, and inductive generalization", In G.E. Hinton & J. A. Anderson (eds.), *Parallel models of associative memory*. Hillsdale, NJ: Erlbaum.