The Effects of Pattern Presentation on Interference in Backpropagation Networks

Jacob M.J. Murre[1]
Unit of Experimental and Theoretical Psychology
Leiden University

Abstract

This paper reviews six approaches to solving the problem of 'catastrophic sequential interference'. It is concluded that all of these methods function by reducing (or circumventing) hidden-layer overlap. A new method is presented, called 'random rehearsal training', that further explores an approach introduced by Hetherington and Seidenberg (1989). A constant number of patterns, randomly selected from those learned earlier, is rehearsed with every newly learned pattern. This scheme of rehearsing patterns may, perhaps, be compared to the functioning of the 'articulatory loop' (Baddeley, 1986). It is shown that this presentation method may virtually eliminate sequential interference.

Preventing 'Catastrophic Interference'

Both from a psychological and from a practical point of view, standard backpropagation models (Rumelhart, Hinton, and Williams, 1986) suffer from an important weakness: on sequential learning tasks they exhibit strong retroactive interference. Newly learned patterns may erase nearly all existing memories (Grossberg, 1987; McCloskey and Cohen, 1989; Ratcliff, 1990). This behavioral implausibility has become the subject of many studies, usually with reference to the name 'catastrophic interference' coined by McCloskey and Cohen (1989). Several proposals have been made to overcome the strong interference in sequential learning tasks.

A number of studies approaches the issue by enhancing the network architecture or the learning rule. French (1991), for example, uses a method whereby after prolonged learning only k nodes in the hidden layer remain active for each pattern. He calls this method 'k-node sharpening'. Kortge (1990) proposes a new learning rule, called the 'novelty rule'. With this rule, the amount of learning is made dependent on the relative novelty of the input pattern. Sloman and Rumelhart (in press) use a network without hidden units, and with weights that are logically gated by 'episodic units' (i.e., representing the learning context). It seems that networks without hidden units, in general, are less prone to sequential interference (Lewandowsky, 1991; Hetherington, 1990b). Hinton and Plaut (1987) use a network in which the hidden units have either 'fast' or 'slow' weights. Since they focused primarily on the effects of retraining items earlier in the list, this approach does not directly address the problem as posed by either McCloskey and Cohen (1989) or Ratcliff (1990). We might, finally, mention the model by Kruschke (1992) in which the receptive fields of the hidden-layer units are functionally located (restricted) before learning.

Apart from an alteration of the working of the backpropagation algorithm, interference may be reduced by merely changing the representation of the input and output patterns. Some studies have successfully used bipolar pattern features (i.e., values -0.5 and 0.5, or values in this range; Kortge, 1990; Lewandowsky, 1991). Others have argued that normalization of the pattern length may reduce interference (Kruschke, personal communication). It may also be noted that the nature of the patterns used seems to have a strong effect on sequential interference. For example, Brouse and Smolensky (1989) and Hetherington (1990b) have argued that in combinatorial domains (i.e., with a large number of structured patterns, such as words) interference is strongly reduced. Also, Hetherington (1990a) has pointed out that with auto-associative learning one may expect less interference than with hetero-associative learning (i.e., inputs differ from outputs).

As a third general approach we may distinguish between variations in the method of pattern presentation. Hetherington and Seidenberg (1989) trained a network in overlapping blocks (see Table 1, and below), which greatly reduced sequential interference. In this paper, we will focus on another method of presentation, called 'random rehearsal', that is akin to their method. After a brief review of the simulations by McCloskey and Cohen (1989) and Hetherington and Seidenberg (1989), we will describe this new method. In the Discussion, we shall argue that all successful approaches to reducing interference are based on a single underlying factor:

[1] The author is presently employed at the MRC Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF, UK. E-mail: Jaap.Murre@MRC-APU.CAM.AC.UK.

orthogonalization of hidden-layer representations across subsequent patterns or pattern blocks.

Pattern Presentation and Interference

The study by McCloskey and Cohen (1989) aimed at teaching (by 'rote learning') a model some simple arithmetic: adding, subtracting, dividing, and multiplying numbers in the range zero to nine. During training, two numbers and an arithmetic operator were presented as input patterns, while the correct answer was presented as output. The model consisted of a straightforward, three-layer backpropagation model with fully connected layers. Numbers (single digits) were represented by activating three consecutive nodes in the output or input layer. The number three, for example, was represented as 0 0 0 1 1 1 0 0 0 0 ....., and a zero as 1 1 1 0 0 0 0 ..... The input layer consisted of 28 input nodes (two times twelve nodes for representing the digits, and four additional nodes for the operators), the hidden layer consisted of fifty nodes, and the output layer consisted of 24 nodes (twelve for digits and twelve for tens).

The network could easily be taught all summed digit pairs, as well as all multiplied digit pairs. The pairs were presented for training in blocks with varying random order. When the network was trained on patterns drawn from the entire training set, no problems occurred. But when the network was first taught all additions with one (e.g., [1+1], [2+1], [3+1], ..., and also [1+2], [1+3], [1+4], ...), and only *then* on all additions with two (except [1+2] and [2+1], which had already been learned), the newly learned patterns appeared to have washed out all memory of addition with one. Performance on the ones, dropped from 100% to 57%, after a single run, and to 30%, after two runs on the twos.

The simulations by McCloskey and Cohen (1989) were replicated by Hetherington and Seidenberg (1989) with essentially similar results. A second simulation by these authors, however, indicated that learning the twos does not completely destroy all memory of the ones. It appeared that the ones were relearned faster than a totally novel set of additions (with three). The model thus showed evidence of 'savings' (Ebbinghaus, 1985): the ones were not completely unlearned, which greatly accelerated relearning. Based on these results (also see Hinton and Plaut, 1987), Hetherington and Seidenberg (1989) argue that the catastrophic interference found by McCloskey and Cohen (1989) is primarily dependent on the method of pattern presentation. In particular, they argue that *blocking* of learning trials (i.e., first a block of ones, then a block of twos) may be an important contributing factor, and that "if, instead of following this strict blocking scheme, there is some minimal retraining on the ones, performance will rapidly improve due to savings." (p.30) Based on this idea they used a training scheme intermediate between both strict blocking and fully concurrent presentation of patterns.

Stage	Pattern sets
1.	1 1
2.	1 1 2
3.	1 2 2 3
4.	1 2 2 3 3 4
5.	2 3 3 4 4 5

Table 1. *Training scheme used by Hetherington and Seidenberg (1989). The table shows the sets presented in each stage. A stage lasted for ten epochs. In each epoch, patterns were presented in a different random order.*

Hetherington and Seidenberg (1989) trained their model in five stages on addition with ones, twos, threes, and fours. For each addition, a set was constructed containing thirteen digit pairs as mentioned above. The sets were constructed, so that they would not overlap (i.e., [1+3] occurred in either the set of ones, or the set of threes, but not in both). The training scheme is shown in Table 1. Presenting two sets of one type corresponds to presenting each element of the set twice, in random order, interleaved with elements from other sets. From the table it becomes clear that the consecutive stages overlap: patterns are trained during a number of consecutive stages. At stage five, the ones are no longer retrained, so that on the basis of the above cited data we might expect considerable interference as a result of training the twos, threes, fours, and fives, while not simultaneously retraining ones. The results, however, indicate that this is not the case. After training on stage 5, the model is still able to correctly reproduce in between twelve and thirteen ones (out of a possible thirteen). The authors further report that after continued training for 35 more epochs following stage four, the mean number of correct responses on the ones was still 91% (11.8 out of 13). Their conclusion, therefore, is that this method of pattern presentation prevents catastrophic interference.

We did a series of simulations to investigate further the effects of pattern presentation schemes on retroactive interference. Our findings indicate that Hetherington and Seidenberg's (1989) results on the detrimental effects of strict blocking do not directly generalize to other models. A method similar to their 'method of overlapping stages', however, appears to work well on

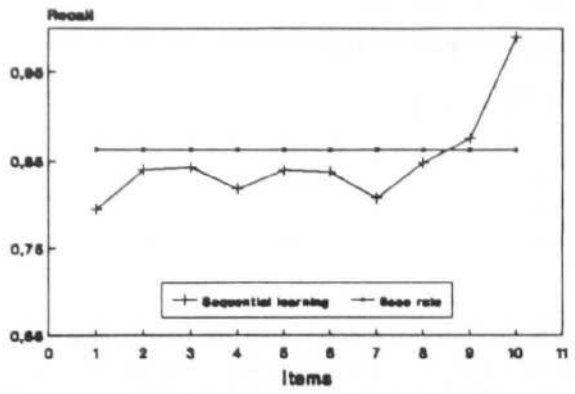

Fig.1. *Interference in backpropagation as a result of strict sequential learning. The results are averaged over 100 replications.*

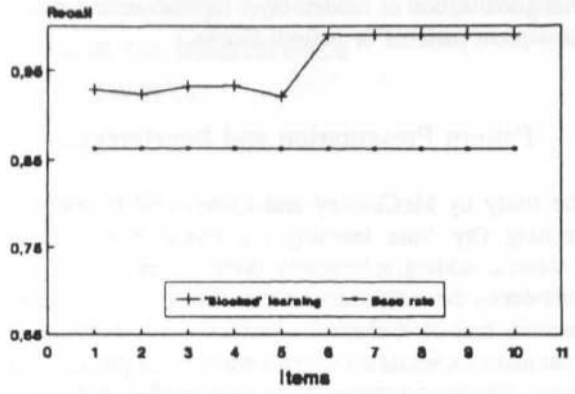

Fig.2. *Interference in backpropagation as a result of a strict blocking scheme. The results are averaged over 100 replications.*

auto-associative learning of random pattern vectors. The patterns used in our simulations consisted of eight elements. Ten new patterns were generated for each simulation (and replication). Each of the pattern elements was assigned a (uniform) random value between zero and one. The length of the vector was normalized to 1.0 (this may have reduced sequential interference, such as between blocks, see discussion below). The model used was a simple three-layer backpropagation network. The size of the input and output layers was eight, the size of the hidden layer was five nodes (simulations indicated that increasing the hidden layer beyond this size did not essentially influence the results, see Figure 6 and the discussion below). Before every simulation, weights were (uniform) randomly initialized in the range [-0.5, 0.5]. The learning rate was 0.5, the momentum parameter was set at 0.9. With these parameters, the networks easily learns ten random patterns to the criterion described below.

Simulation 1. In this simulation, 'strict sequential learning' was used. Each of the patterns was learned until the criterion was reached, and *not repeated thereafter*. The criterion consisted of a correlation coefficient (i.e., cosine of the angle between the two vectors) of more than 0.99 between the (target) pattern and pattern produced at the output layer. The simulation was repeated 100 times. For each replicated simulation both the initial weights and the patterns were generated anew. The averaged results are shown in Figure 1. Recall is represented by the correlations remaining after having learned all patterns. The base rate shown in the figure is the expected correlation of 0.863 between a random pattern and its output before the network has learned anything. It was established by generating 5000 random patterns and averaging the correlations. As can be seen from Figure 1, strict sequential learning causes catastrophic interference to the extent that after learning the network performs actually *worse* than before learning.

Simulation 2. Having established that in this simulation paradigm strict sequential learning gives rise to 'more than catastrophic interference', we trained the network using 'strict blocking' of trials. First, five patterns were simultaneously trained until the criterion was reached (see above), followed by training on patterns six to ten. After these had reached the criterion, the network was tested for recall. The simulation was repeated 100 times. The results are shown in Figure 2. Strict blocking also leads to considerable retroactive interference, although not as bad as strict sequential learning.

Simulation 3. To test whether training in 'overlapping stages', as described by Hetherington and Seidenberg (1989) is a feasible method for reducing interference the following training method was used. A fixed-size 'window' was moved over the ordered pattern set. All patterns in the window were trained to the set criterion. Say, the window can contain three patterns (we will speak of a *depth* of 3). Then, the training stages are as follows. In subsequent stages we train patterns A, B, C, .., as follows: (A), (A,B), (A,B,C); (B,C,D); (C,D,E), etc. Simulations were carried out with windows of depth 1, 2, 3, and 4. Network, patterns, and parameters were as above. For each depth, 100 replications were carried out. The averaged results are shown in Figure 3. A depth of 1 leads to strong retroactive interference, comparable to using zero depth (i.e., strict sequential learning). Depths of 2, 3, and 4, however, lead a to progressively decreasing interference, although even a depth of 4 performs hardly better than strictly blocked learning in this respect.

Simulation 4. According to Hetherington and Seidenberg (1989), the overlapping stages method leads to reduced interference, because old patterns are occasionally retrained. The 'windowing method' of the previous simulations only rehearses the most recent

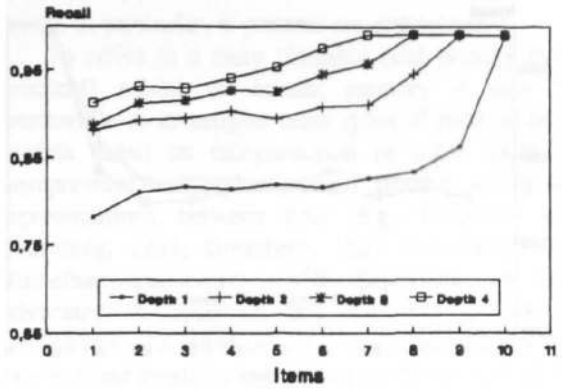

Fig.3. *Interference in backpropagation using a windowed training method. Each point is averaged over 100 replications.*

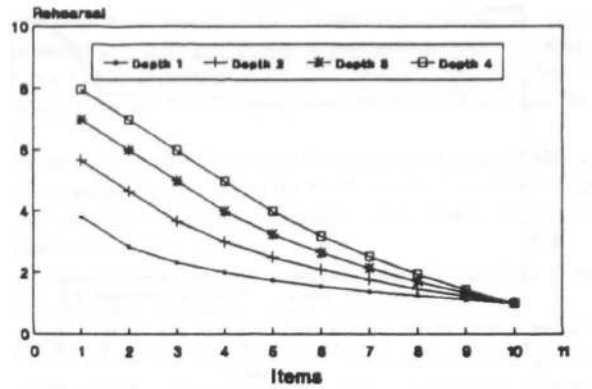

Fig.4. *Expected number of item rehearsals (out of a possible 10) for depths of 1 to 4. See text for an explanation.*

patterns. It would be interesting to see whether an increase in performance (i.e., a reduction in interference) could be achieved by rehearsing a constant number of patterns randomly chosen from the already learned patterns. In the method used, the first patterns have a higher chance of being rehearsed than late patterns in the list. Exact chances of rehearsal with list size 10 are shown in Figure 4. With a depth of two, for example, pattern 3 will on the average be rehearsed about four times (out of a possible ten). The term *depth* is maintained, although here it refers to randomly selected items. We remark, furthermore, that if depth is 3, it implies that the first four items will certainly be rehearsed up until pattern 4.

The results are given in Figure 5. As can be seen, the 'random rehearsal method' works successfully. Especially with depths of 3 and 4 retroactive interference is strongly reduced. Note, that the total number of rehearsals is not greater than that of the windowed training scheme.

Discussion

Simulations 1 to 4 convincingly demonstrate that pattern presentation schemes may considerably influence retroactive interference, from 'more than catastrophic' for strictly sequential learning to 'only slightly' for the random rehearsal method.

As was argued by Hetherington and Seidenberg (1989), their method of presentation may be called plausible from a psychological point of view. Occasional reminders of items are nearly always given during prolonged instruction (e.g., in classroom situations). In our method of 'random rehearsal', reminders are drawn from all items learned, rather that just the most recent ones. This may or may not be a plausible scheme for classroom instruction. We would rather argue, however, that it can be viewed as a partial implementation of the 'articulatory loop' proposed by Baddeley and Hitch (1974; Baddeley, 1986). While learning a list, a fixed number of items is drawn from memory and rehearsed with the new items. As is shown in Figure 4, earlier items have a high chance of being retrained. The articulatory loop may thus be seen as a method whereby occasional 'reminders' are generated for retraining (which occurs quickly due to savings, see above). The proposed method is only a partial implementation, because it presuposes that all items learned are still fully available. In a more complete implementation, rehearsal should be a more self-contained process. Use could, for instance, be made of a recurrent scheme to implement the articulatory loop. Such studies have indeed been carried out recently (Burgess and Hitch, 1991; Nolfi, Parisi, Vallar, and Burani, 1990; Mul, Phaf, and Wolters, in preparation). In these models, early items in the list are recalled *better* than late items. In a recurrent network an 'articulatory loop' may, thus, fully counteract the effects of retrograde interference. More importantly, perhaps, is that this primacy effect is consistent with well-established experimental findings.

One fact that remains surprising in the simulations presented above, is that a seemingly important architectural characteristic such as the size of the hidden layer has a negligible influence on retroactive interference. In Figure 6, for example, a replication of Simulation 1 is shown with hidden layers of 5 and 20 nodes. As can be seen, increasing the size of the hidden layer has only a minor influence on retroactive interference (this was also found by Hetherington and Seidenberg, 1989, see their note 3, p.32). In fact, increasing the hidden layer seems to have a slight *negative* effect on recall.

This effect may be explained with reference to the

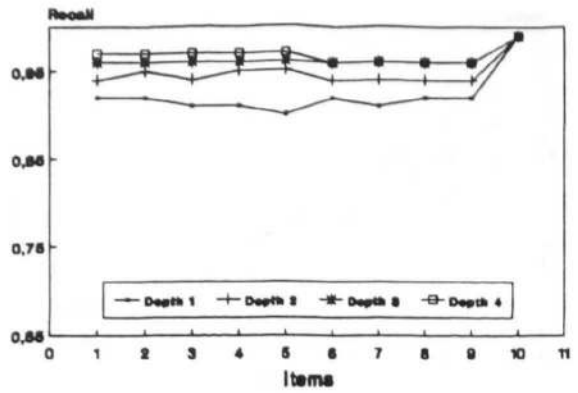

Fig.5. *Interference in backpropagation using a random rehearsal method. Each point is averaged over 100 replications.*

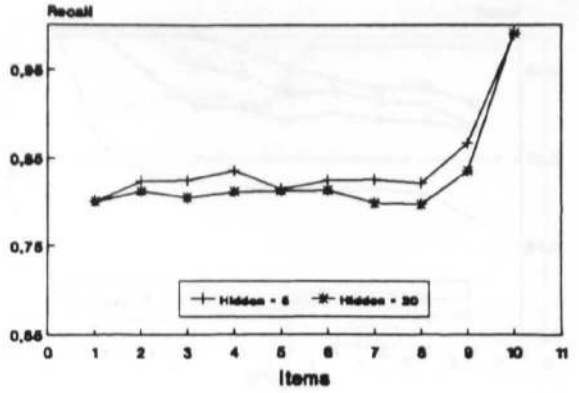

Fig.6. *The negligible effect of increasing the size of the hidden layer on retroactive interference. The results are averaged over 100 replications.*

nature of hidden-layer representations emerging in standard backpropagation networks. A detailed analysis of Ratcliff's (1990) first series of simulations has revealed that nearly all sequential interference can be attributed to overlap of hidden-layer representations (Murre, 1992). Backpropagation is able to develop sufficiently distinct representations for patterns *within* a list (block). The representations *between* lists, however, are almost purely random. It can be shown that as few as two overlapping active hidden units (i.e., in two sequentially learned patterns) may cause nearly complete unlearning of the first pattern (Murre, 1992). Normally, roughly half of the hidden-layer will be active for any pattern presented. Chances of an overlap of more than one active unit are, therefore, very high. This is even more so, if the size of the hidden-layer is increased, which may explain the effect of Figure 6.

Reducing hidden-layer overlap will decrease sequential interference (also see French, 1991). In fact, all studies cited above that succeed in reducing interference share this as a common factor:

1. ***Sharpening*** (French, 1990). This method is introduced explicitly to reduce hidden-layer overlap. Fewer active nodes results in a lower chance of overlap. A more detailed analysis of this method, however, shows that in many situations there is a rather high chance that learning will not converge, because within-list representations may overlap (Murre, 1992).
2. ***Novelty rule*** (Kortge, 1990). This method results in emphasizing the between-pattern differences, which gives rise to more distinct hidden-layer representations.
3. ***Restricted receptive fields*** (Kruschke, 1992). Functionally located hidden nodes only respond to a restricted part of the pattern space. This results in a decreased overlap of representations: a unit that responds to a certain pattern, is less likely to respond to another.
4. ***Normalizing patterns*** (Kruschke, personal communication). Normalization results in more restricted receptive fields. This can be most easily seen when considering binary nodes. Suppose, that a certain hidden-unit has weights [1,1] and threshold 1.3. Such a node will be activated by, for example, the pattern (0.8,0.6), but also by all patterns (>1.3,>1.3). Normalization places all patterns on a hypersphere. After normalization, the unit will still be activated by pattern (0.8,0.6), which already has length one. Patterns in the area of (2,0), (0,3), etc., however, will be normalized to (1,0), (0,1), etc., so that the unit is no longer activated by such patterns. Similarly, a linear-sigmoid node can be placed so that it 'carves off' just a small piece of the hypersphere. With normalized patterns such nodes develop more restricted receptive fields, which gives rise to more distinct hidden-layer representations.
5. ***Bipolar pattern features*** (Kortge 1990; Lewandowsky, 1991). With bipolar pattern features, it can easily be shown that for orthogonal patterns, hidden units tend to develop orthogonal representations (also between lists, which is not the case with standard backpropagation, see Ratcliff, 1990).
6. ***Windowed training*** (Hetherington and Seidenberg, 1989). Windowed training is based on providing reminders of recent items. This enables the algorithm to develop representations that are orthogonal over several subsequent lists rather than just within-lists.
7. ***Random rehearsal training*** (Simulation 4, above). This method functions similarly to windowed training. It also provides reminders of items learned earlier. These reminders enable the model to keep representations orthogonal.

We conclude from this brief review that all of these methods derive their succes primarily from making the hidden-layer representations of patterns between-lists more orthogonal. One way of circumventing the problem would be to have a model without hidden layers. This has indeed been found (Hetherington, 1990b, Lewandowsky, 1991, and Sloman and Rumelhart, in

press), in particular, if patterns are orthogonal.

To arrive at a more plausible (and perhaps more practical) model of human memory it may be worthwhile to investigate other types of models. Most models based on categorization or other forms of unsupervised learning are able to develop orthogonal representations between lists (e.g., Carpenter and Grossberg, 1987; Grossberg, 1987; Kohonen, 1990; Rumelhart and Zipser, 1985). Elsewhere, we have advocated an approach that combines a modular architecture with intramodular competition. The learning rate in these modules is sensitive to the novelty of the incoming pattern (Murre, Phaf, and Wolters, 1989, 1992; Murre, 1992). This approach combines several partial solutions to the problem of sequential interference outlined above. Once this problem has been fully understood, we may be able to perform detailed simulations of a more challenging phenomenon: the forgetting gradient in retrograde amnesia. Severe disturbance to the brain seems to result in a loss of recently learned patterns, with patterns learned earlier being saved (Squire, 1987). This well-established fact seems to run counter to any neural network model devised thus far.

References

Baddeley, A.D. (1986). *Working memory*. Oxford: Clarendon Press.

Baddeley, A.D., & G.J. Hitch (1974). Working memory. In: G.H. Bower (Ed.) *Recent advances in the psychology of learning and motivation, Vol.VIII*. New York: Academic Press, 47-90.

Brouse, O., & P. Smolensky (1989). Virtual memories and massive generalization in connectionist combinatorial learning. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 380-387.

Burgess, N., & G. Hitch (1991). Towards a network model of the articulatory loop. *Journal of Memory and Language*, in press.

Carpenter, G.A., & S. Grossberg (1987). Neural dynamics of category learning and recognition: attention, memory consolidation, and amnesia. In: J. Davis, R. Newburgh & E. Wegman (Eds.) *Brain structure, learning, and memory*. AAAS Symposium Series.

Ebbinghaus, H. (1885). *Über das Gedächtnis*. Leipzig: Dunker.

French, R.M. (1991). Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 173-178.

Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science, 11*, 23-63.

Hetherington, P.A. (1990a). Interference and generalization in connectionist networks: within-domain structure or between-domain correlation? *Neural Network Review*, 4, 27-29.

Hetherington, P.A. (1990b). *The sequential learning problem in connectionist networks*. Unpublished Master's thesis, Department of Psychology, McGill University, Montreal, Canada.

Hetherington, P.A., & M.S. Seidenberg (1989). Is there 'catastrophic interference' in connectionist networks? *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 26-33.

Hinton, G.E., & D.C. Plaut (1987). Using fast weights to deblur old memories. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 177-186.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78, 1464-1480.

Kortge, C.A. (1990). Episodic memory in connectionist networks. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 764-771.

Kruschke, J.K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44.

Lewandowsky, S. (1991). Gradual unlearning and catastrophic interference: a comparison of distributed architectures. In: W.E. Hockley & S. Lewandowsky (Eds.) *Relating theory and data: essays on human memory in honour of Bennet B. Murdoch*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

McCloskey, M., & N.J. Cohen (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In: G.H. Bower (Ed.) *The psychology of learning and motivation*. New York: Academic Press.

Mul, N.M., R.H. Phaf, and G. Wolters (in prep.). Sequential recurrent networks in CALM: rehearsal, short-term memory, and ordered recall. Internal Report, Unit of Experimental and Theoretical Psychology, Leiden University.

Murre, J.M.J. (1992). *Categorization and learning in modular neural networks*. Hemel Hempstead: Harvester Wheatsheaf, forthcoming in September 1992.

Murre, J.M.J., R.H. Phaf, & G. Wolters (1989). CALM networks: a modular approach to supervised and unsupervised learning. *Proceedings of the International Joint Conference on Neural Networks Washington DC, 1*, 649-656. (New York: IEEE Press. IEEE Catalog Number 89CH2765-6).

Murre, J.M.J., R.H. Phaf, G. Wolters (1992). CALM: Categorizing And Learning Module. *Neural Networks, 5*, 55-82.

Nolfi, S., D. Parisi, G. Vallar, & C. Burani (1990). Recall of sequences of items by a neural network. In: D.S. Touretzky, J.L. Elman, T.J. Sejnowski, & G.E. Hinton (Eds.) *Proceedings of the 1990 Connectionist Summer School*. San Matteo CA: Morgan Kaufmann.

Ratcliff, R. (1990). Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review, 97*, 285-308.

Rumelhart, D.E., G.E. Hinton, & R.J. Williams (1986). Learning internal representations by error propagation. In: D.E. Rumelhart & J.L. McClelland (Eds.) *Parallel distributed processing. Volume 1: Foundations*. Cambridge, MA: MIT Press.

Rumelhart, D.E., & D. Zipser (1985). Feature discovery by competitive learning. *Cognitive Science, 9*, 75-112.

Sloman, S.A., & D.E. Rumelhart (in press). Reducing interference in distributed memory through episodic gating. In: A. Healy, S. Kosslyn, & R. Shiffrin (Eds.) *Essays in honor of W.K. Estes*.

Squire, L.R. (1987). *Memory and brain*. Oxford: Oxford University Press.