Catastrophic Interference is Eliminated in Pretrained Networks

Ken McRae, University of Rochester
and
Phil A. Hetherington, McGill University

When modeling strictly sequential experimental memory tasks, such as serial list learning, connectionist networks appear to experience excessive retroactive interference, known as catastrophic interference (McCloskey & Cohen,1989; Ratcliff, 1990). The main cause of this interference is overlap among representations at the hidden unit layer (French, 1991; Hetherington,1991; Murre, 1992). This can be alleviated by constraining the number of hidden units allocated to representing each item, thus reducing overlap and interference (French, 1991; Kruschke, 1992). When human subjects perform a laboratory memory experiment, they arrive with a wealth of prior knowledge that is relevant to performing the task. If a network is given the benefit of relevant prior knowledge, the representation of new items is constrained *naturally*, so that a sequential task involving novel items can be performed with little interference. Three laboratory memory experiments (ABA free recall, serial list, and ABA paired-associate learning) are used to show that little or no interference is found in networks that have been pretrained with a simple and relevant knowledge base. Thus, catastrophic interference is eliminated when critical aspects of simulations are made to be more analogous to the corresponding human situation.

Learning in standard, feed forward, back propagation networks (hereafter, standard networks) is a result of altering weights on connections between units via feedback or experience. Because patterns are represented in a distributed manner, many units and weights participate in encoding an item, enabling hidden units to capture meaningful relations among items. Because items are superimposed during learning, retroactive interference may occur; events occurring later in the training regime result in poor performance on previously-learned items (McCloskey & Cohen, 1989; Ratcliff, 1990). Although this is also true for humans, McCloskey and Cohen claim that interference in these networks is "catastrophic"; learning on later trials results in grossly impaired performance on previously learned items.

The sequential learning problem was demonstrated by Ratcliff (1990) in an attempt to model serial list learning in humans. In this task, a subject is sequentially presented with a number of items and is asked to recall them after the final item has been presented. In Ratcliff's simulation, 16 items were trained individually and sequentially. The network retained only the final item; performance on items 1 to 15 was very poor. Manipulations of various parameters, such as learning rate and momentum, did not significantly improve network performance. Similar failures to resolve the sequential learning problem by manipulating network parameters was reported by McCloskey and Cohen (1989).

In this article, we outline the major cause of catastrophic interference in standard networks, describe recent approaches to the problem, and present a novel approach. In contrast to previous work on the sequential learning problem that has manipulated network parameters and architecture (e.g., Hinton & Plaut, 1987; Kortge, 1990; Kruschke, 1992; Lewandowsky, 1991; Sloman & Rumelhart, 1992) we attempt to make the simulation more similar to the human situation that is being modeled. When this is done, catastrophic interference is eliminated.

Overlap at the Hidden Unit Layer

Standard networks tend to unlearn catastrophically because there are too few constraints on hidden unit representations when learning initial items (French, 1991; Hetherington, 1991; Kruschke, 1992; Murre, 1992). Even completely non-overlapping stimulus patterns can overlap at the hidden unit layer (Hetherington, 1991). Most or all hidden units encode initial items, so later learning necessarily involves changing weights that encode previous items. If, however, a network was constrained to allocate a limited subset of hidden units to encode initial items, then overlap between old and new items would be reduced, decreasing interference.

There have recently been a number of proposals advanced to impose limits on the number of hidden units used to encode early items. In a standard network, the receptive field of a hidden unit includes

all input units. Thus, each hidden unit has equal probability of encoding two input patterns, maximizing the potential for interference. Kruschke (1992) used hidden units with local receptive fields to reduce overlap among items at the hidden unit layer. If a hidden unit is connected to only a subset of the input units, then the probability that two items are encoded by the same hidden unit decreases. In a variation of Kruschke's (1992) approach, French (1991) selectively "sharpened" a subset of the hidden units so that some became more active than others during initial learning, and fewer were used to encode each early item. Interference was reduced when a subset of hidden units were sharpened.

A Natural Solution

Catastrophic interference has been found in simulations of strictly sequential verbal learning experiments. Subjects in these experiments are typically college students who enter into an experiment possessing a large body of knowledge about words and their properties. If a network is pretrained so that it also possesses a body of knowledge prior to commencement of a sequential learning simulation, the first item (or short list of items) in a sequential learning task must be learned within the constraints resulting from prior knowledge. Many of the hidden units are already committed to representing only a limited aspect of the input space by virtue of strong weights from a limited number of input units. When a second item (or short list) is trained, interference with the first is decreased because the probability of overlap at the hidden unit layer is reduced. Thus, in accord with French's (1991) and Kruschke's (1992) ideas, our proposal also focuses on the state of the hidden unit layer at commencement of sequential training, but in contrast, we take advantage of the fact that training on a corpus of items *naturally* constrains hidden units' receptive fields and sharpens their activations.

A Small, Naive Model

The first simulation is a simple demonstration of the effect of pretraining on the number of hidden units used to encode a pattern. An encoder network was used, with 8 input and output units, and 5 hidden units. Six distributed patterns were used. The first pattern was represented by turning on the first three units and setting the rest to zero, /11100000/, the second pattern by the successive three units, /01110000/, and so forth. The initial weights were randomly selected from a uniform distribution, within the range ±.5. Training involved repeatedly presenting the fourth pattern until error fell below 0.01, at which point, hidden unit activations were recorded. Figure 1 shows a typical run, in which, after training pattern 4, each hidden unit contributed approximately equally to the output. Hence, as suggested above, without constraints on hidden unit activations, a standard network encodes a pattern across a large subset of its hidden units.

Figure 1. Representation of a single pattern across the hidden unit layer in a small, naive network. The pattern is distributed across all hidden units.

A Small, Pretrained Model

The network was configured as in the previous simulation. Pretraining consisted of training all patterns except pattern 4 until mean error fell below 0.01. Pattern 4 was then trained to the same criterion. Figure 2 shows that only three hidden units participated in its encoding. Thus, like the constraints introduced by French (1991) and Kruschke (1992), pretraining reduced the number of hidden units used to encode an item.

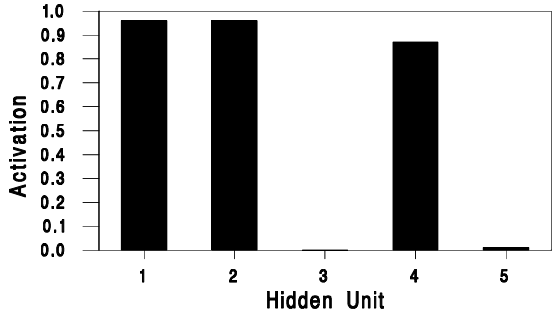

Figure 2: Representation of a single pattern across the hidden unit layer in a small, pretrained network. The pattern was encoded using only 3 of 5 hidden units.

This demonstration implies that if a network is constrained by pretraining it, then hidden units respond selectively to new input. Suppose that a

network is made more analogous to adult humans by providing it with previous knowledge before subjecting it to a strictly sequential training regime associated with a recognition memory or serial list learning experiment. Given the demonstration above, previous training may enable a network to perform with little interference in strictly sequential learning tasks. To test this prediction, larger simulations of free recall, serial-list learning, and paired-associate memory tasks were conducted. They are presented below.

Pretraining a Larger Network

For three memory tasks (ABA free-recall, ABA paired-associate, serial-list), performance of naive and pretrained networks was compared. Networks were pretrained either on the orthographic representations of the set of 2897 English monosyllabic words used in the Seidenberg and McClelland (1989) model of naming, or on 1448 orthographic stimulus-response pairs constructed from words in that corpus. The input and output patterns were 400-unit wickelgraph representations of orthography (see Seidenberg & McClelland). A wickelgraph representation is a distributed representation of word spelling that encodes letter order and item similarity. For the free-recall and serial-list learning tasks, an encoder network was used; hence, input and output patterns were identical. Pretraining on the 2897 words continued for 44 epochs, until mean error fell below 10. At this point, the network correctly reproduced all words; the output for a word matched itself better than any of the other 2896. For the pattern associator, stimuli consisted of every second word (alphabetically) from the Seidenberg and McClelland corpus. A response was chosen randomly from among the 2897 items, with the constraint that no response occurred greater than three times. The network was pretrained for 40 epochs, at which point mean error was below 13 and it computed the correct response to each of the 1448 stimuli. In summary, in each simulation, a pretrained network that possessed knowledge of English orthography was compared to a naive network that was a tabula rasa.

For each simulation, the degree of hidden unit overlap is reported, followed by network performance on the list learning task. We compared the degree of hidden unit overlap between pairs of CVCs using a measure of percentage overlap ([1 - sum of absolute differences/sum of total activation] * 100). According to this measure, identical representations score 100, and non-overlapping representations score 0. Analyses of network performance will be described in each section.

Free-Recall ABA

In an ABA free-recall memory task, a subject memorizes a list of items, such as consonant-vowel-consonant strings (list A), followed by a second list (list B). She is then asked to recall first list items in any order. An encoder network (400 input/output, 150 hidden units) was used to simulate this task.

The stimuli were consonant-vowel-consonant trigrams (CVCs) that were medium similarity items taken from a typical verbal learning experiment (Underwood, 1952). None of them were words from the pretraining corpus. The CVCs were randomly mixed to create five replications of A and B lists, each containing eight CVCs. During CVC training in this and the following two simulations, the momentum parameter in the McClelland and Rumelhart (1988) simulation software package, was set to zero. The first list was trained until mean error fell to approximately 18. At this point in training, the output for each pattern was closer to the target CVC than any of the other 2204 (21*5*20) possible CVCs. However, they were not as well learned as the pretrained words; CVCs learned in an experiment would not be as ingrained as a college student's knowledge of common English words.

Overlap

Similar to the smal simulations, the distributions of hidden unit activations were compared for sets of CVCs in the naive and pretrained networks. Figure 3 shows a histogram of hidden unit activations for YUG after it was trained to criterion along with seven other CVCs (list A). Note that the representation of YUG in the naive network was distributed fairly evenly across all hidden units. In contrast, in the pretrained network, almost all were inactive, with 38 of 150 fully activated. Thus, the naive network spread the representation of YUG across its hidden units, but the pretrained network used only a small subset. These results were not unique to YUG: the distribution of hidden unit activations averaged across CVCs was approximately uniform in the naive network, but bimodal in the pretrained case.

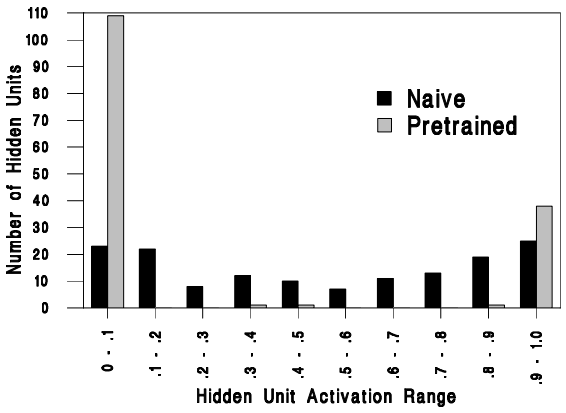

Figure 3. Distribution of hidden unit activation in representation of YUG in naive versus pretrained networks. Almost all hidden units encoded YUG in the naive network, but only a small minority in the pretrained network.

In the naive network, percentage overlap between JEP and YUG was 58%, and 82% between JEP and ZEP. In contrast, in the pretrained model, percentage overlap between JEP and YUG was 24%, and 58% between JEP and ZEP. The mean hidden unit overlap between pairs of CVCs in the naive model (65%, *SD* = 6%) was significantly greater than in the pretrained model (36%, *SD* = 8%), $t(27) = 34.34$, $p < .0001$. Note, also, that as well as reducing overlap, pretraining preserved similarity information (i.e., the hidden unit representation of JEP was more similar to ZEP than to YUG).

Interference

Given that a network with encoded knowledge of English represented CVCs with less overlap, interference in a free-recall ABA simulation should be drastically reduced. After the first list of 8 CVCs was trained, error was recorded for each pattern in both lists as a second 8-item list was trained (list B). Because the second list was learned at different rates in the naive and pretrained networks, it was inappropriate to directly contrast amount of interference at each recorded epoch. To compare interference in the naive and pretrained networks, we required a measure of interference on first-list items given the degree to which second-list items had been learned. Therefore, linear regression analyses were conducted in which total error for second-list items was used to predict total error for first-list items. The slope of the regression line indicates the rate at which first-list error increased relative to the decrease in second-list error (i.e., first-list interference in terms of second-list learning). In both the naive and pretrained cases, second-list error predicted first-list error (naive: $R^2 = .97$; pretrained: $R^2 = .91$). Critically, significantly less interference obtained when a network possessed knowledge of English orthography. The 95% confidence intervals for the slope of the function relating second-list learning to first-list retention were computed (naive: -.34 to -.26; pretrained: -.10 to -..06). The slope for the naive network was significantly steeper, and the confidence intervals were non overlapping. In other words, for each increment in second-list learning, first-list error increased significantly more in a naive network. Interference can also be compared at the point at which output for second-list items was closer to the true response than to any other CVC. At the point where the second list was learned, total error had risen 69.43 (a 47% increase) for the first-list items in the naive network, but only 0.21 (a 0% increase) for those same items in the pretrained network. In summary, when an encoder network possessed knowledge of English orthography prior to an ABA free-recall simulation, catastrophic interference was eliminated.

Serial-List Learning

The identical network and stimuli were used to simulate serial-list learning. In a free-recall serial-list learning task, subjects are presented with a number of items sequentially, and are asked to recall them in any order.

Overlap

Analogous to the previous simulation, YUG, when trained individually until its error fell below 18, was represented over the hidden units in an approximately uniform manner in the naive network, but by only 38 units in the pretrained network. Mean overlap between CVC pairs was significantly greater in the naive (65%, *SD* = 3%) than in the pretrained network (37%, *SD* = 9%), $t(119) = 35.03$, $p < .0001$. For example, when YUG was learned, followed by NOL, their representations overlapped by 60% in the naive network, compared to 29% in its pretrained counterpart. Thus, in terms of retention of YUG, there was greater potential for interference in the naive network. Interestingly, pretraining also preserved the role of item similarity in modulating interference. To illustrate, FAX and FAH, two highly similar CVCs, overlapped 68% in the naive network, barely above mean overlap (65%). In contrast, they

overlapped 71% in the pretrained network, well above the mean (37%). Thus, although similarity between items plays little or no role in a network with an unconstrained hidden unit layer (see Hetherington, 1991), it plays a role when new information must be incorporated into an existing knowledge base. Thus, while the potential for interference has almost been eliminated, generalization has not (see Ratcliff, 1990).

Interference

To simulate serial list learning, five orders of the 16 CVCs were prepared. For each replication, they were sequentially trained to criterion (error less than 18). Following training on the final pattern, all CVCs were tested for retention. Error by serial position, averaged across the five replications, is displayed in Figure 4. With the naive network, a replication of Ratcliff's (1990) results obtained; retention was very poor for all but the final item. In contrast, all patterns were well retained when pretraining was provided. The performance of the naive and pretrained networks was compared using a t-test. Across serial position, interference in the pretrained network was significantly less than in the naive network, $t(15) = 13.08$, $p < .0001$. In fact, it may be that performance of the pretrained network was too good; there was less interference than is typically found in human serial-list learning experiments.

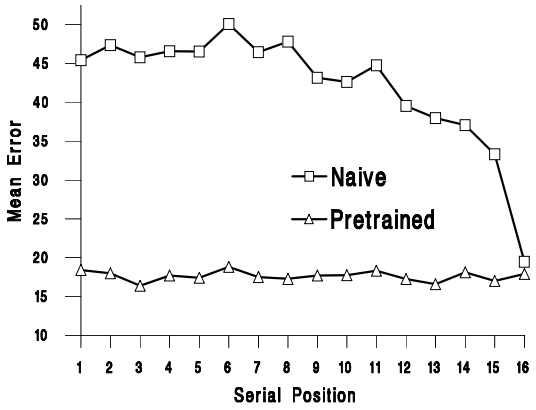

Figure 4. Mean performance of naive versus pretrained networks in the serial list-learning task. The upper (naive) and lower (pretrained) lines represent performance after all items were trained.

Paired-Associate Learning

In an ABA paired-associated task, a subject learns a list of stimulus-response pairs (list A), followed by a second list (list B), and then is asked to recall the first list. The importance of conducting both free-recall and paired-associate simulations was illustrated by Hetherington (1990, 1991). A factor such as item similarity that decreases interference in the simulation of one task may increase it in another. Thus, the ABA pattern association task was simulated to demonstrate that the effect of pretraining on reduction of interference is independent of task (i.e., type of network).

The paired-associate ABA simulation was conducted in the same manner as its free-recall counterpart. However, because the network learned arbitrary associations, a greater number of hidden units were required; 600 hidden units were used to encode 1448 word pairs. Sixteen random CVC-CVC stimulus-response pairs were created for the ABA simulation. They were randomly mixed to create five pairs of A and B lists, each containing eight patterns. For each replication, the first list was trained until mean error fell to approximately 18.

Overlap

The distributions of hidden unit activations for items in this simulation were approximately uniform in the naive networks, but bimodal in the pretrained network. BAB-BIX, for example, activated only nine hidden units greater than .9 in the pretrained network. Mean overlap between CVC pairs was significantly greater in the naive (62%, $SD = 2\%$) than in the pretrained network (19%, $SD = 6\%$), $t(27) = 35.90$, $p < .0001$.

Interference

During second-list training, error was recorded for patterns in both lists. As in the free-recall ABA simulation, in both the naive and pretrained networks, second-list error predicted first-list error (naive: $R^2 = .91$; pretrained: $R^2 = .86$). Interference was significantly reduced when a network was pretrained. The 95% confidence intervals for the slope of the function relating second-list learning to first-list retention were computed (naive: -.24 to -.15; pretrained: -.08 to -.04). The slope for the naive network was significantly steeper, and the confidence intervals were non overlapping. That is, in a pretrained network, for each increment in second-list learning, first-list error rose significantly less. Interference was also compared at the point at which the output for all second-list items was closer to the target response than to any other CVC. At the point where the second list was learned, error had risen 120.71 (an 82% increase) for the 8 items in the naive network, but only 14.15 (a 9% increase) in the pretrained case. Thus, when a standard network

possessed knowledge of English orthography prior to a paired-associate ABA simulation, interference was drastically reduced. Again, interference may been reduced to below what is found with humans in paired associate ABA experiments.

Summary and Conclusion

The three previous simulations demonstrated that interference was drastically reduced in simulations of verbal learning experiments using a simple and principled manipulation. College students typically serve as subjects in these experiments, and enter the experimental situation possessing an impressive body of knowledge that can be used to perform a task. When networks were given the benefit of prior relevant knowledge, catastrophic interference disappeared. Despite the absence of interference, similar items still overlapped at the hidden unit layer, so generalization should still occur. However, there was less interference in the pretrained networks than is typically found with humans in analogous verbal learning experiments.

The simulations reported here are not sophisticated models of human memory experiments; in fact, they are insufficient implementations of each task. Our purpose was not to capture the variety of behavioral effects seen in these tasks, but merely to use the simulations to demonstrate that catastrophic interference in standard networks can be alleviated using a simple and principled manipulation. The choice of specific networks and procedures was primarily motivated by a desire to deviate little from the simulations of McCloskey and Cohen (1989) and Ratcliff (1990). It is left for future work to produce more sophisticated and realistic simulations of ABA free-recall, ABA paired-associate, and serial-list learning tasks.

References

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

Hetherington, P.A. (1990). Interference and generalization in connectionist networks: Within-domain structure or between-domain correlation? [Review of O. Brousse & P. Smolensky (1989). *Virtual memories and massive generalization in connectionist combinatorial learning.*] *Neural Network Review, 4*(1), 27-29.

Hetherington, P.A. (1991). *The Sequential Learning Problem in Connectionist Networks*. Unpublished Master's Thesis, McGill University, Montreal, Quebec.

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

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

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

Lewandowsky, S. (1991). Gradual unlearning and catastrophic interference: A comparison of distributed architectures. To appear in W.E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock..* Hillsdale, NJ: Erlbaum.

McClelland, J. L., & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. Cambridge MA.: MIT Press.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The catastrophic interference. In G.H. Bower (Ed.), *The psychology of learning and motivation: Volume 24* (pp. 109-164). New York: Academic Press.

Murre, J.M.J. (1992). *Learning and categorization in modular neural networks*. Hillsdale, NJ: Lawrence Erlbaum.

Ratcliff, R. (1990). Connectionist Models of Recognition Memory: Constraints Imposed by Learning and Forgetting Functions. *Psychological Review, 97,* 285-308.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed developmental model of word recognition and naming. *Psychological Review, 96*, 523-568.

Sloman, S. A., & Rumelhart, D. E. (1992). Reducing interference in distributed memory through episodic gating. In A.F. Healey, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to cognitive processes: Essays in honor of William K. Estes.* Hillsdale, NJ: Erlbaum.