French, R. M. (2003) Catastrophic Forgetting in Connectionist Networks. In Nadel, L. (Ed.) *Encyclopedia of Cognitive Science. Vol. 1*, pp. 431 - 435. London: Nature Publishing Group.

# Catastrophic interference in connectionist networks

Robert M. French
Quantitative Psychology and Cognitive Science
Department of Psychology
University of Liège,
4000 Liège, Belgium
Tel. (32.4) 366.20.10
FAX: (32.4) 366.28.59
email: rfrench@ulg.ac.be
URL: http://www.ulg.ac.be/cogsci/rfrench.html

## I. Headword

Catastrophic interference in connectionist networks

## II. Keywords

Catastrophic interference, catastrophic forgetting, forgetting in neural networks, connectionist representations.

## III. Contents list

## IV. Article definition

Unlike human brains, connectionist networks can forget previously learning information suddenly and completely (i.e., "catastrophically") when learning new information. Various solutions for overcoming this problem are discussed.

## V. Introduction

The connectionist paradigm in artificial intelligence burst into prominence in 1986 with the publication of Rumelhart and McClelland's two-volume collection of articles entitled *Parallel Distributed Processing: Explorations into the Micro-structure of Cognition.* The "connectionist revolution," as it is sometimes called today, unquestionably began with the publication of this book. Some twenty years earlier, research on an elementary type of neural network, perceptrons (forerunners of modern connectionist networks), had come to a sudden halt in 1969 with the publication of *Perceptrons,* Minsky and Papert's careful mathematical analysis of the capacities of a particular class of single-layered perceptrons. Minsky and Papert's work demonstrated a number of fundamental theoretical limitations of elementary perceptrons. Multi-layered perceptrons and new learning algorithms were developed over the course of the next two decades that allowed these limitations to be overcome. These new networks were able to do many of the things that posed severe problems for traditional symbolic artificial intelligence programs. For example, they were able to function appropriately with degraded inputs, they could generalize well, they were fault-tolerant, etc. The late 1980's marked a gold-rush period of attempts to apply these networks to everything from underwater mine detection to cognition, from stock market prediction to bank-loan screening.

At the end of the 1980's, however, a problem with these multi-layered perceptrons networks came to light. McCloskey & Cohen (1989) and Ratcliff (1990) showed that the very property — namely, a single set of weights that served as the network's memory — that gave these networks such power was the root cause of an unsuspected problem: catastrophic interference. Grossberg (1982) had previously cast this problem in the more general context of "stability-plasticity." In short, the problem was to determine how network architectures could be designed that would be simultaneously sensitive to new input without being overly disrupted by it?

Catastrophic interference is the "stability-plasticity" problem in spades. It occurs when a network has learned to recognize a particular set of patterns and then is called upon to learn a new set of patterns. The learning of the new patterns modifies the weights of the network in such a way that the originally learned set of patterns is forgotten. In other words, the newly learned patterns suddenly and completely — "catastrophically" — erase the network's memory of the previously learned patterns.

## VI. Main Text

### Catastrophic forgetting vs. normal forgetting

Catastrophic forgetting is significantly different from normal human forgetting. The latter is a normal cognitive process; the former is basically unknown in human cogntition. Some of the best known experiments in human forgetting were conducted by Barnes and Underwood (1959). Subjects begin by learning a set of paired associates (A-B) consisting of a nonword and a real word (e.g., *pruth – heavy*, etc.). Once this learning is complete, they learn to associate a new real word with each of the original nonwords (A-C). At various points during the learning of the A-C pairs, they asked subjects to recall the originally learned A-B associates. McCloskey and Cohen (1989) ran a similar A-B/A-C paradigm using addition facts on a standard connectionist network. After five learning trials in the A-C condition, the network's knowledge of the A-B pairs had dropped to 1% and was completely gone after 15 trials. In other words, the newly learned pairs had catastrophically interfered with the originally learned associated pairs.

The problem for connectionist models of human memory — in particular, those models with a single set of shared multiplicative weights (e.g., feedforward backpropagation networks) — is that catastrophic interference, for all intents and purposes, is not observed in humans. This raises a number of issues of significant practical and theoretical interest. Arguably, the most important issue for cognitive science is understanding how the brain managed to overcome the problem of catastrophic forgetting. The brain is, after all, a distributed (or semi-distributed) neural network, and yet, does not exhibit anything like the catastrophic interference seen in connectionist networks. In particular, what neural architecture allowed the brain to overcome catastrophic interference, and what characteristics of neural networks, in general, will allow them to overcome this?

At present, in order to avoid catastrophic interference, most connectionist architectures rely on learning algorithms which require the network to cycle through all of the patterns to be learned over and over, gradually adjusting the weights a small amount with each pattern. Finally, after many cycles (called *epochs*) through the entire set of patterns, the network will (usually) converge on an appropriate set of weights for the set of patterns that it is supposed to learn. The problem is that humans do not learn in this way remotely similar to this. In order to memorize ten piano pieces, say, we do not play each piece once and then cycle through all the pieces over and over until we have learned them all. We learn piano pieces — and just about everything else — *sequentially*. In other words, we start by learning one or two pieces thoroughly and then learn a new piece, then another, and so on. However, were a standard connectionist network to do this, each new piece learned by the network would most likely completely erase from its memory all previously learned pieces. Certainly, by the tenth piece the network would have no recollection whatsoever of the first piece. The moral of the story is that, in order for connectionist networks to exhibit anything like human sequential learning, they must overcome the problem of catastrophic interference.

## Measures of catastrophic interference

In this section we will define the two most common measures of catastrophic interference — namely, "exact recognition" and "relearning." In both cases, the network is trained to criterion on an initial set of patterns. It is then given a second set of patterns to learn. Once it has learned this second set of patterns, we use one of the two measures of forgetting to determine the effect on the originally learned patterns of having learned the second set of patterns. For the exact-recognition measure, we check to see what percentage of the original patterns can still be recognized by the network. (In other words, we give the input part of the pattern to the network and see what output it produces. If it produces the correct output, the network is considered to have "recognized" the pattern.) The relearning measure, first proposed by Ebbinghaus for human memory in the late 19th century, involves seeing how long it takes the network to relearn the originally learned patterns. Thus, even if the rate of exact recognition is very low, the knowledge might lie "just below the surface" and be able to be relearned very quickly. The original studies by McCloskey & Cohen (1989) and Ratcliff (1990) relied on an exact-recognition measure. A study by Hetherington and Seidenberg (1989) using the relearning measure showed that, at least in some cases, catastrophic interference might be less of a problem than was thought because the network, even if it could not recognize the originally learned patterns exactly, it could relearn them very quickly.

## Solutions to the problem

Attempts to solve the problem came quickly. One of the first was the suggestion by Kortge (1990) that the problem was due to the backpropagation learning algorithm. He proposed a modified learning algorithm using what he called "novelty vectors" that did, in fact, decrease catastrophic interference. The key idea of novelty vectors was "blame assignment." Kortge's learning rule was developed for auto-associative networks, i.e., networks that, starting from a random weight configuration, learn to produce on output the vectors that they received on input (hence the name "auto-associator" for this type of network). Each pattern to be learned was fed through the network and the output was compared with the intended output (i.e., the input). Kortge called this difference vector a "novelty vector" because the bigger the activation differences at each node, the more novel the input, since for vectors that the network had already learned there would be little difference between output and input. This novelty vector indicated where to change the weights the most: the greater the novelty activation, the more the weights were changed. This technique did, indeed, significantly reduce catastrophic interference, even though the technique only applied to auto-associative networks.

French (1992) argued that catastrophic forgetting was in large measure due to excessive overlap of *internal* representations. He claimed that the problem lay with the fully distributed nature of the network's internal representations and suggested that by developing algorithms that produced "semi-distributed" internal representations (i.e. representations whose activation was spread only over a limited set of total number of hidden nodes) catastrophic interference could be reduced. To this end he suggested a learning algorithm, "node sharpening," that developed far sparser internal representations than standard backpropagation. The result was a significant reduction in catastrophic interference. The overly sparse representations developed by this technique, however, resulted in a significant decrease in the network's ability to discriminate categories. What was needed was a means of making representations as highly distributed as possible and, at the same, time as separated as possible.

Brousse and Smolensky (1989) and McRae and Hetherington (1993) showed that the problem was closely related to the domain of learning. In domains with a high degree of internal structure, such as language learning, the problem is much less acute. McRae and Hetherington (1993) eliminated the problem by pre-training the network on a random sample of patterns drawn from the domain. Because of the pre-existing structure in the domain, this sample was enough to capture the overall regularities in the domain. Consequently, the new patterns to be learned were perceived by the network to be nothing more than variants of already-learned patterns and did not interfere with previously learning.

The early attempts to solve the problem of catastrophic interference attempted to reduce representational overlap on input or internally. Kortge (1990) and Lewandowsky (1991) modified the input vectors in an attempt to achieve greater mutual orthogonalization (this is equivalent to reducing the overlap among input vectors). French (1992), Murre (1992), Krushke (1993) and others developed algorithms that reduced internal representational overlap and, in doing so, managed to significantly reduce the amount of catastrophic interference.

Certain authors (e.g., Carpenter, 1994) have laid the blame for the problem of catastrophic interference on a particular architectural feature of the most widely used class of connectionist networks, namely their use of multiplicative connection-weights. In the ART family of networks (Carpenter & Grossberg, 1987), new input does not interfere with previously learned patterns because the network is able to recognize new patterns as being new and assigns a new set of nodes for their internal representation.

Hopfield networks, and related architectures, have been shown to have critical saturation limits beyond which there is a radical drop off of memory performance. For these networks, unlearning is initially gradual and then, after the memory becomes saturated, forgetting becomes catastrophic.

## Rehearsal and pseudorehearsal

Most connectionist networks learn patterns concurrently, which, in terms of human cognition, is a very contrived type of learning. For a given set of *n* patterns, $\{P_1, ..., P_n\}$, the network will successively adjust its weights by a very small amount for all of the patterns and then will repeat this process until the network has found an appropriate set of weights that allow it to recognize all *n* patterns. This, in itself, is a strange and non-cognitive way of learning. In addition, if a new set of patterns $\{P_{n+1}, ..., P_m\}$ must then be learned by the network, the standard way of handling the situation is to go find the original set of patterns, mix them in with the new patterns to be learned, creating a new set $\{P_1, ..., P_n, P_{n+1}, ..., P_m\}$ and then train the network on this new expanded set. In this way, the new patterns will indeed not interfere with the old patterns, but there is a major problem with this technique — namely, in the real world, *the originally learned patterns are often no longer available* and cannot simply be added to the set of new patterns to be learned.

In 1995 Anthony Robins made a major contribution to research on catastrophic forgetting with a technique based on what he called *pseudopatterns* (Robins, 1995). His idea was simple and elegant. Suppose that a connectionist network with *n* inputs and *m* outputs has learned a number of input-output patterns $\{P_1, P_2, . . ., P_N\}$ generated by some underlying function *f*. Assume that these original input-output vectors are no longer available. How could one determine, even approximately, what function the network had originally learned? One way would be to create a number, *M*, of random input vectors of length *n*, $\{\hat{\imath}_1, ..., \hat{\imath}_M\}$. These pseudo-input vectors would be fed through the previously trained network, producing a set of outputs $\{\hat{o}_1, ..., \hat{o}_M\}$ corresponding to each of the pseudo-inputs. This would result in a set of *pseudopatterns*: $S = \{\psi_1, \psi_2, . . ., \psi_M\}$ where $\psi_1$: $\hat{\imath}_1 \rightarrow \hat{o}_{1;}$ $\psi_2$: $\hat{\imath}_2 \rightarrow \hat{o}_{2;\ . . .}$ $\psi_N$: $\hat{\imath}_M \rightarrow \hat{o}_M$. This set of pseudopatterns would approximate the prior learning of the network. The accuracy of the pseudopatterns in describing the originally learned function would depend on the nature of the originally learned function. Thus, when the network had to learn a new set of patterns to be learned, it would mix in a number of pseudopatterns with the new patterns to be learned.

The pseudopattern technique was the basis of dual-memory models developed by French (1997) and Ans & Rousset (1997) which loosely simulate the hippocampal-neocortical separation, considered by some to be the brain's way of overcoming catastrophic interference (McClelland, McNaughton, O'Reilly, 1995). These models incorporate two separate, continually interacting pattern-processing areas, one for early-processing, one for long-term storage, information being passed back and forth between the areas by means of pseudopatterns. This allows them to forget gradually and to perform sequential learning appropriately. Somewhat unexpectedly, these dual-memory networks also exhibit a gradual representational "compression" (i.e., fewer active nodes) over time of the long-term internal representations, a fact which, if it can be shown to also occur in humans, might help explain certain types of category-specific deficits commonly observed in amnesiacs (French & Mareschal, 1998).

## Other techniques for alleviating catastrophic forgetting in neural networks

A number of other techniques have been developed to address the problem of catastrophic interference. Notably, there have been attempts to combine auto-associative architectures with sparse representations to reduce the level of catastrophic interference. Architectures using two different kinds of weights on the connections between nodes, one which decays rapidly to zero, the other that decays much more slowly. Convolution-correlation models such as CHARM and TODAM, which are mathematically equivalent to certain types of connectionist networks (sigma-pi networks) seem to be relatively immune to catastrophic interference, at least up to a point. Cascade-correlation learning algorithms have also been tried as a means of alleviating catastrophic interference with some success. For a more complete review of the various models that have been developed to handle the problem of catastrophic interference in connectionist networks, see French (1999).

**Summary**

The problem of catastrophic interference in connectionist networks has been known and studied since the early 1990's. The problem is of particular importance because sequential learning of the kind done by humans cannot be achieved unless a solution is found to this problem. In other words, network models of cognition must, as Grossberg has stressed, be sensitive to new input but not so sensitive that the new input destroys previously learned information. Certain types of patterns, such as those found in highly structured domains, are less susceptible to catastrophic interference than patterns from less well structured domains. Nature seems to have evolved a way of keeping new learning (hippocampal learning) at arms' length from previously learned information stored in the neo-cortex (neo-cortical consolidation), thus physically preventing new learning from interfering with previously learned information. Connectionist models have been developed that simulate this cerebral separation. This may not be — in fact, is certainly not — the only way to route to solving the problem of catastrophic interference, but its close relationship with the way in which the brain may have solved the problem, makes further exploration of these dual-memory models of particular interest.

## VII. References

Ans, B. and Rousset, S. (1997) Avoiding catastrophic forgetting by coupling two reverberating neural networks. *Academie des Sciences, Sciences de la vie*, 320, 989 - 997

Barnes, J. and Underwood, B. (1959) "Fate" of first-learned associations in transfer theory. *Journal of Experimental Psychology, 58*, 97-105

Brousse, O. and Smolensky, P. (1989) Virtual Memories and Massive Generalization in Connectionist Combinatorial Learning. In *Proceedings of the 11th Annual Conference of the Cognitive Science Society,* 380-387, NJ:LEA

Carpenter, G. (1994) A distributed outstar network for spatial pattern learning. *Neural Networks*, *7*, 159-168

Carpenter, G. and Grossberg, S. (1987) A massively parallel archetecture for a self-organizing neural pattern recognition machine." *Computer Vision, Graphics and Image Processing, 37*, 54-115

French, R. M. (1992) Semi-distributed Representations and Catastrophic Forgetting in Connectionist Networks, *Connection Science, 4,* 365-377

French, R. M. (1997) Pseudo-recurrent connectionist networks: An approach to the "sensitivity–stability" dilemma. *Connection Science, 9*, 353-379

French, R. M. (1999). Catastrophic Forgetting in Connectionist Networks. *Trends in Cognitive Sciences, 3*(4), 128-135.

French, R. M. and Mareschal, D. (1998) Could category-specific anomia reflect differences in the distributions of features within a unified semantic memory? In

*Proceedings of the Twentieth Annual Cognitive Science Society Conference*, 374-379, NJ:LEA.

Grossberg. S. (1982) *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control.* Boston: Reidel

Hetherington, P. and Seidenberg, M., (1989), Is there “catastrophic interference” in connectionist networks?, In *Proceedings of the 11th Annual Conference of the Cognitive Science Society*, 26-33, Hillsdale, NJ: LEA

Kortge, C., (1990) Episodic Memory in Connectionist Networks. *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, 764-771, Hillsdale, NJ: LEA

Krushke, J. (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, in W. Hockley and S. Lewandowsky (eds.) *Relating Theory and Data: Essays on Human Memory in Honor of Bennet B. Murdock*, 445-476, NJ: LEA

McClelland,J., McNaughton,B. and O'Reilly,R. (1995) Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102, 419-457

McCloskey, M. and Cohen, N. (1989) Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (ed.) *The Psychology of Learning and Motivation: Vol. 24*, 109-164, NY: Academic Press

McRae, K. and Hetherington, P. (1993) Catastrophic interference is eliminated in pretrained networks. In *Proceedings of the 15h Annual Conference of the Cognitive Science Society*, 723-728, Hillsdale, NJ: LEA

Murre, J. (1992) *Learning and Categorization in Modular Neural Networks*. Hillsdale, NJ: LEA

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

Robins, A. (1995) Catastrophic forgetting, rehearsal, and pseudorehearsal. *Connection Science*, 7, 123 – 146

## VIII. Bibliography

French, R. M. (1999). Catastrophic Forgetting in Connectionist Networks. *Trends in Cognitive Sciences, 3*(4), 128-135.

Hetherington, P. (1991) *The Sequential Learning Problem.* Master’s Thesis. Department of Psychology, McGill University, Montreal, Québec, Canada

## IX. Glossary

Forgetting, catastrophic forgetting, catastrophic interference, interference in neural networks, dual-network memory, hippocampus-neocortex interaction, memory consolidation, pseudorehearsal, overlap of representations, connectionist network learning algorithms.

## X. Illustrations/tables

None

## XI. Cross references to other articles

--

## XIII. Information on word processing package used

Word 97 for Windows