Catastrophic forgetting in connectionist networks

Robert M. French

All natural cognitive systems, and, in particular, our own, gradually forget previously learned information. Plausible models of human cognition should therefore exhibit similar patterns of gradual forgetting of old information as new information is acquired. Only rarely does new learning in natural cognitive systems completely disrupt or erase previously learned information; that is, natural cognitive systems do not, in general, forget 'catastrophically'. Unfortunately, though, catastrophic forgetting does occur under certain circumstances in distributed connectionist networks. The very features that give these networks their remarkable abilities to generalize, to function in the presence of degraded input, and so on, are found to be the root cause of catastrophic forgetting. The challenge in this field is to discover how to keep the advantages of distributed connectionist networks while avoiding the problem of catastrophic forgetting. In this article the causes, consequences and numerous solutions to the problem of catastrophic forgetting in neural networks are examined. The review will consider how the brain might have overcome this problem and will also explore the consequences of this solution for distributed connectionist networks.

By the end of the 1980s many of the early problems with connectionist networks, such as their difficulties with sequence-learning and the profoundly stimulus–response nature of supervised learning algorithms such as error backpropagation had been largely solved. However, as these problems were being solved, another was discovered by McCloskey and Cohen[1] and Ratcliff[2]. They suggested that there might be a fundamental limitation to this type of distributed architecture, in the same way that Minsky and Papert[3] had shown twenty years before that there were certain fundamental limitations to what a perceptron[4,5] could do. They observed that under certain conditions, the process of learning a new set of patterns suddenly and completely erased a network's knowledge of what it had already learned. They referred to this phenomenon as catastrophic interference (or catastrophic forgetting) and suggested that the underlying reason for this difficulty was the very thing – a single set of shared weights – that gave the networks their remarkable abilities to generalize and degrade gracefully.

Catastrophic interference is a radical manifestation of a more general problem for connectionist models of memory – in fact, for any model of memory – the so-called 'stability–plasticity' problem[6,7]. The problem is how to design a system that is simultaneously sensitive to, but not radically disrupted by, new input.

This article will focus primarily on a particular, widely used class of distributed neural network architectures – namely, those with a single set of shared (or partially shared) multiplicative weights. While this defines a very broad class of networks, this definition is certainly not exhaustive. The remainder of this article will discuss the numerous attempts over the last decade to solve this problem within the context of this type of network.

Catastrophic interference versus gradual interference

First, we need to make clear the distinction between what McCloskey and Cohen[1] call 'the mere occurrence of interference' and 'catastrophic interference.' Barnes and Underwood[8] conducted a series of experiments that measured the extent of retroactive interference in human learning. They first had subjects learn a set of paired associates (A–B) consisting of a nonsense syllable and an adjective (e.g. *dax* with *regal*, etc.) and then asked them to learn a new set of paired associates (A–C) consisting of the same nonsense syllables associated with a new set of adjectives. They were able to determine that the forgetting curve for the A–B associate pairs produced by interference from the learning of the new A–C pairs was relatively gradual. By contrast, McCloskey and Cohen[1] showed that, at least under certain circumstances, forgetting in a standard backpropagation network was anything but gradual. In one set of experiments, a standard backpropagation network[9] thoroughly learned a set of 'one's addition facts' (i.e. the 17 sums 1+1 through 9+1 and 1+2 through 1+9). Then the network

R.M. French is at the Quantitative Psychology and Cognitive Science Unit, Department of Psychology, University of Liège, 4000 Liège, Belgium.

tel: +32 4 221 05 42
fax: +32 4 366 28 59
e-mail: rfrench@ulg.ac.be
http://www.fapse.ulg.ac.be/Lab/Trav/rfrench.html

1364-6613/99/$ – see front matter PII: S1364-6613(99)01294-2

learned the 17 'two's addition facts' (i.e. 2+1 through 2+9 and 1+2 through 9+2). Recall performance on the originally learned 'one's addition facts' plummeted as soon as the network began learning the new 'two's addition facts'. Within 1–5 learning trials of the two's addition facts, the number of correct responses on the one's addition facts had dropped from 100% to 20%. By five more learning trials, this percentage had dropped to 1%, and by 15 trials, no correct answers from the previous one's addition problems could be produced by the network. The network had 'catastrophically' forgotten its one's addition sums. In a subsequent experiment that attempted to more closely match the original Barnes and Underwood paradigm, they again found the same catastrophic, rather than gradual, forgetting in the neural network they tested. Ratcliff[2] tested a series of error-backpropagation models on a number of similar tasks, for vectors of different sizes and for networks of various types, and also found that well-learned information can be catastrophically forgotten by new learning.

These two papers are generally given credit for bringing the problem of catastrophic interference to the attention of the connectionist community. One might wonder why, if the problem of catastrophic interference was as serious as these authors claimed, it had taken more than five years to come to light. McCloskey and Cohen[1] answered this as follows: 'Disruption of old knowledge by new learning is a recognized feature of connectionist models with distributed representations… However, the interference is sometimes described as if it were mild and/or readily avoided… Perhaps for this reason, the interference phenomenon has received surprisingly little attention…' (p. 110).

The conclusions of these two papers raised a number of important theoretical as well as practical concerns – namely: is this problem inherent in *all* distributed architectures? Can distributed architectures be designed that avoid the problem? If human brains are anything like connectionist models, why is there no evidence of this kind of forgetting in humans? Can this kind of forgetting be observed in animals with a less highly evolved brain organization? (see Box 1). And finally, will distributed connectionist networks remain unable to perform true sequential learning?[14] In other words, humans tend to learn one pattern, then another, then another, and so on, and even though some of the earlier patterns may be seen again, this is not necessary for them to be retained in memory. As new patterns are learned, forgetting of old, unrepeated patterns occurs gradually over time. However, for any network subject to catastrophic interference, learning cannot occur in this manner, because the new learning will effectively erase previous learning.

Measuring catastrophic interference

The two initial studies on catastrophic interference[1,2] relied on an 'exact recognition' measure of forgetting. In other words, after the network had learned a set of binary patterns and was then trained on a second set of patterns, its recognition performance on the first set was tested by presenting each old input pattern to the network and seeing how close it came to its originally learned associate. If all of the output nodes were not within 0.5 of the original associate (i.e. could not correctly generalize to the original associate), then the network was said to have 'forgotten' the original pattern. Hetherington and Seidenberg[15] introduced a 'savings' measure of forgetting based on a relearning measure first proposed by Ebbinghaus[16]. To measure how completely the network had lost the original associations, they measured the amount of time it required to relearn the original data. They showed that it was often the case that a network that seems, on the basis of exact-recognition criterion, to have completely forgotten its originally learned associations, can be retrained very quickly to recall those associations. Unfortunately, later work showed that not all catastrophic forgetting is of this 'shallow' sort. Most discussions of catastrophic forgetting now include both of these measures (see also Box 2).

Early attempts to solve the problem

As early as 1990, various solutions were suggested to the problem of catastrophic forgetting. Kortge[17] claimed that the problem was not one inherent in distributed connectionist architectures, but rather was due to the backpropagation learning rule. He developed a variation of the backpropagation algorithm using what he called 'novelty vectors' that produced a decrease in catastrophic interference. The idea is that 'when the network makes an error, we would like to blame just those active units which were 'responsible' for the error – blaming any others leads to excess interference with other patterns' output.' When a new pattern was to be learned by his auto-associative network, it was fed through the network, which produced some pattern on output. The difference between this pattern and the intended output (i.e. the pattern itself, since the task of the network was to produce on output what it had seen on input) was what he called a novelty vector (the bigger the differences, the more 'novel' the pattern). His new weight-change algorithm weighted the standard backpropagation delta parameter based on activation values from this novelty vector. The bigger the novelty activation, the bigger the corresponding weight change.

The effect of Kortge's learning rule was to reduce the amount of overlap between input representations of the new patterns to be learned and previously learned patterns. French[18,19] suggested that, in general, catastrophic forgetting was largely a consequence of the overlap of *internal* distributed representations and that reducing this overlap would reduce catastrophic interference. He argued for the necessity of 'semi-distributed' representations that would remain distributed enough to retain many of the advantages of fully distributed representations, but were not so fully distributed as to overlap with all other representations and cause catastrophic interference. Explicitly decreasing representational overlap by creating 'sparse vectors' (i.e. internal representations in which only a few nodes were active, and most were not active) served as the basis for French's activation sharpening algorithm[18,19]. An extra step is added to the standard backpropagation learning algorithm in which activations patterns at the hidden layer are 'sharpened', that is, the activation level of the most active hidden node(s) is increased slightly for each pattern, while the activations of other nodes are decreased. This technique had the effect of 'sparsifying' the hidden-layer representations and significantly

Box 1. Catastrophic interference in animals

In humans, new learning interferes with old, but the old information is forgotten gradually, rather than catastrophically (Ref. a). McClelland, McNaughton and O'Reilly suggest that this may be due to our hippocampal-neocortical separation (Ref. b). But does catastrophic interference affect other animals and under what conditions? One likely candidate seems to be the learning – and catastrophic forgetting – of information related to time in the rat.

In their natural environment, animals are very capable predictors of important events like periodical food availability, which plausibly involves an ability to represent time. In the laboratory, researchers have developed several techniques to study timing processes. In the 'peak procedure' (Ref. c) rats learn to press a lever to receive food after a certain fixed duration. During each trial, the rate of lever-presses/second is recorded. The moment of maximum lever-pressing is called the 'peak time' and reflects the moment at which the animal maximally expects the food.

The observation of steady-state behavior following training has long been used to understand the mechanisms underlying timing abilities (Refs d–f). Recently it has also been used to study the acquisition of a new temporal representation (Refs g,h; A. Ferrera, PhD thesis, University of Liège, 1998).

We will now compare two scenarios. In the first sequential learning experiment, the animal will first learn a 40-second duration and then an 8-second duration. Once the new 8-second duration is learned, the criterion is switched back to the original 40-second duration. In both cases, the learning of the new duration can be described in terms of a moving peak time. Crucially, the second transition is no faster than the first. In short, there is no evidences of savings from the initial 40-second learning. One reasonable interpretation of this result is that new 8-second learning completely (catastrophically) wiped out the original 40-second learning.

However, things are very different in the concurrent-learning scenario in which the animal learns a 40-second and an 8-second duration *concurrently*. Sometimes food is presented 8 seconds after the start of the trial, sometimes after 40-seconds. The animal is then switched to a 40-second-only reinforcement schedule, which is continued until the animal consistently produces a single peak time of 40 seconds. The reinforcement duration is then switched to 8 seconds and then switched back to 40 seconds again. Unlike the previous case in which there was no savings from its previous learning, the animal, having learned the two durations concurrently, can now rapidly shift back to the correct 8-second duration and, later, to the correct 40-second duration. In this case, while there is forgetting, there is no catastrophic forgetting of the originally learned 8-second duration. This would imply that the representations developed by the rat during concurrent learning are significantly different from those developed during sequential time–duration learning. This is almost precisely what would occur if a simple feedforward backpropagation network were used to model these time-learning data.

McClelland, McNaughton and O'Reilly (Ref. b) suggest that catastrophic forgetting may be avoided in higher mammals because of their development of a hippocampal–neocortical separation. It is an open question whether lower animals in which this separation is absent would suffer from catastrophic interference produced by the sequential learning of patterns likely to interfere with one another, as the sequential learning of a 40-second duration seemed to interfere with the prior learning of an 8-second duration in the rat.

References

a Barnes, J. and Underwood, B. (1959) 'Fate' of first-learned associations in transfer theory *J. Exp. Psychol.* 58, 97–105

b 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 *Psychol. Rev.* 102, 419–457

c Catania, A. (1970) Reinforcement schedules and psychophysical judgements, in *Theory of Reinforcement Schedules* (Schoenfeld, W., ed.), pp. 1–42, Appleton–Century–Crofts

d Roberts, S. (1981) Isolation of an internal clock *J. Exp. Psychol. Anim. Behav. Process.* 7, 242–268

e Gibbon, J., Church, R. and Meck, W. (1984) Scalar timing in memory, in *Timing and Time Perception* (*Annals of the New York Academy of Sciences, Vol. 423*), pp. 52–77, NY Academy of Sciences

f Cheng, K. and Miceli, P. (1996) Modelling timing performance on the peak procedure *Behav. Process.* 37, 137–156

g Meck, W., Komeily-Zadeh, F. and Church, R. (1984) Two-step acquisition: modification of an internal clock's criterion *J. Exp. Psychol. Anim. Behav. Process.* 10, 297–306

h Lejeune, H. *et al.* (1997) Adjusting to changes in the time of reinforcement: peak interval transition in rats *J. Exp. Psychol. Anim. Behav. Process.* 23, 211–231

reduced catastrophic forgetting as long as there were not too many patterns to be learned.

Brousse and Smolensky[20] acknowledged that catastrophic interference was, indeed, a problem, but they attempted to show that in combinatorially structured domains, such as language and fact learning, neural networks were able to largely overcome the problem. McRae and Hetherington[21] came to a similar conclusion and demonstrated that for domains in which patterns have a high degree of internal structure, when a network is pre-trained on random samples from that domain, catastrophic interference disappears from future learning. The intuitive interpretation of this is that when a domain is quite regular, learning a random sample of exemplars from that domain will be enough to 'capture' the regularities of the domain. Subsequently, when new exemplars are presented to the network, they will tend to be very much like previously-learned exemplars and will not interfere with them. It has also been shown (R. Brander, Master's thesis, University of Queensland, 1998) that in combinatorial domains, especially when sparse internal coding was achieved, there is a significant reduction of catastrophic interference. However, the price paid for the sparse coding in these domains is poorer generalization to new exemplars and poorer overall ability to discriminate[22–24] (see also Box 3).

Reducing representational overlap

In one way or another, almost all of the early techniques relied on reducing representational overlap. Some attempted to use orthogonal recoding of inputs[17,25,26]. These techniques used bipolar feature coding (i.e. –1/1 on each feature input instead of the more standard 0/1 encoding), which made orthogonalization at the input layer easier. One problem with these techniques remains how to determine, in general, how this orthogonal coding on input can be done.

Alternatively, internal representational overlap was reduced by attempting to orthogonalize the hidden-layer activation patterns[18,24,27–31]. It turned out that internal orthogonalization of representations could be made to emerge automatically by pre-training[21]. These models all develop, in one way or another, semi-distributed (i.e. not fully distributed) internal representations within a single network. Because these representations overlap with one another less than fully distributed representations, catastrophic interference is reduced. In some cases, for example, in Sloman and

Box 2. Why the problem of catastrophic interference is so hard

When a distributed network has learns to recognize an initial set of patterns, this means that it has found a point in weight-space, $W_{initial}$, for which the network can recognize all of the patterns it has seen. If the network now learns a new set of patterns, even if the new set is small, it will move to a new solution point in weight-space, W_{new}, corresponding to a set of weights that allows the network to recognize the new patterns. Catastrophic forgetting occurs when the new weight vector is completely inappropriate as a solution for the originally learned patterns (Refs a,b).

Now, if the 'topography' of weight-space were predictable and varied smoothly from one point to the next, catastrophic forgetting would not be a problem. In fact, as McCloskey and Cohen point out, from the outset everyone working with distributed networks knew that new information would adversely affect already-learned information, but no one realized just how bad it would be (Ref. a).

It turns out that weight-space is not exactly a network-friendly place. The very existence of catastrophic forgetting suggested the presence of 'weight cliffs,' that is, areas where moving even small distances over the weight-landscape would radically disrupt prior learning. A paper by Kolen and Pollack clearly confirmed these intuitions (Ref. c). They showed that even extremely small changes in the initial weights of a network could have immense effects on convergence times, even for an extremely simple network (2-2-1) and an extremely simple problem (XOR) (see Fig. I.). It is immediately obvious why this would imply that a move from $W_{initial}$ to W_{new} could have catastrophic effects on the previously learned knowledge, both in terms of the exact-recognition and retraining-time criteria for forgetting.

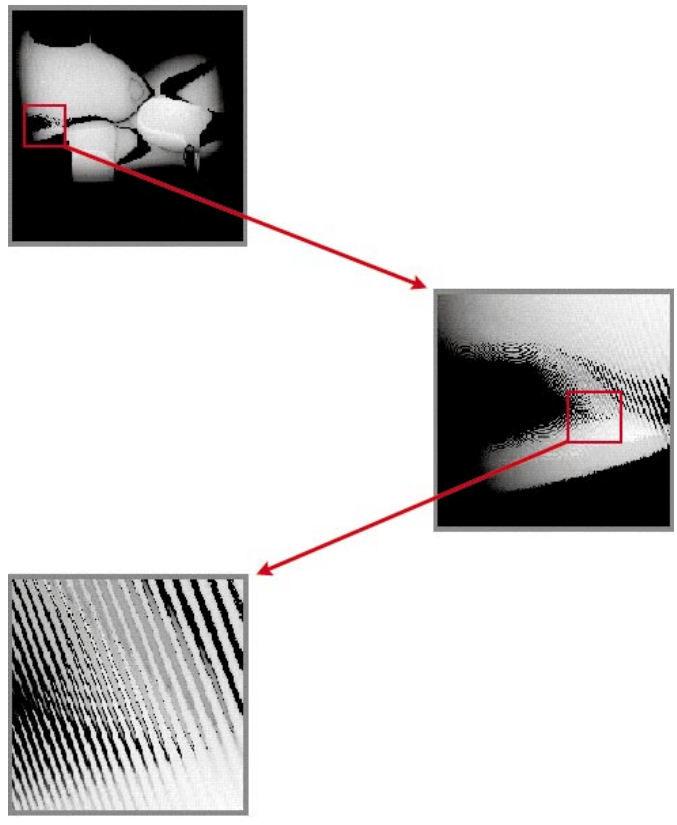


Fig. I. These three panels show variations in convergence times for a 2-2-1 feed-forward backpropagation network learning XOR. Two of the nine weights (i.e. six connection weights and three bias weights) are varied, one along the x-axis, the other along the y-axis. In the top panel, the increments are of size 0.1 and the weights range from –10 to +10. In other words, 40,000 initial weight combinations were examined. The second and third panels also examine 40,000 combinations of the two weights, but 'zoom in' on the area in the square from the previous panel. White indicates the fastest convergence; black indicates no convergence within 200 epochs. The patterns seen show the 'weight cliffs' that represent the unpredictable outcome for convergence of even small movements in weight space for such networks. (Reprinted, with permission, from Ref. c.)

References

a McCloskey, M. and Cohen, N. (1989) Catastrophic interference in connectionist networks: the sequential learning problem, in *The Psychology Of Learning And Motivation* (Vol. 24) (Bower, G. H., ed.), pp. 109–164, Academic Press

b Ratcliff, R. (1990) Connectionist models of recognition memory: constraints imposed by learning and forgetting functions *Psychol. Rev.* 97, 285–308

c Kolen, J. and Pollack, J. (1990) Backpropagation is sensitive to initial conditions *Complex Syst.* 4, 269–280

Rumelhart[31], what amount to localist representations emerge directly from the architecture to avoid the problem. Predictably, in all cases, this reduces catastrophic interference.

Arguably, the two most interesting examples in this latter group are CALM[28] and ALCOVE[29,30]. CALM is an explicitly modular connectionist architecture. A CALM network is made up of three distinct type of competing nodes, excitatory (R-nodes), inhibitory (V-nodes) and arousal nodes (A-nodes). When new input arrives, the system is designed to trigger 'an elaborative search among nodes that mostly have not yet been committed to other representations'[28]. This is reminiscent of the ART family of architectures[7,32] in which a similar notion of representational assignment of new patterns is used. In ART, as in CALM, new input does not, in general, interfere with old input because it is 'recognized' as being new and is 'assigned' to a new node or set of nodes. In other words, a series of 'top-down' connections ensure that only similar patterns are directed to the same node. Once again, new, unfamiliar input is separated from old, familiar patterns, thus allowing catastrophic interference to be avoided. One of the central claims of this line of research is that the key to solving catastrophic forgetting lies with the type of synaptic transfer function used by the model[33]. In this view, the problem of catastrophic forgetting arises from the use of multiplicative path-weights, a very widely accepted part of neural network design. (For a more complete discussion of the role of synaptic transfer functions in catastrophic forgetting in the ART family of networks, see Ref. 33.)

ALCOVE is a three-layer feed-forward network in which the activation of a node in the hidden layer depends (according to an inverse exponential function) on that node's distance from the input stimulus[29,30]. 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 topological localization does not exist in standard backpropagation networks. 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 old representations. This means that the set of weights that produced the old representations will remain largely unaffected by new input.

Box 3. Catastrophic 'remembering'

One of the most complete analyses of the problem of catastrophic interference appeared in Sharkey and Sharkey (Ref. a). In this paper they carefully analysed the underlying reasons for this type of interference and, significantly, introduced the notion of catastrophic remembering. Unlike catastrophic forgetting, catastrophic remembering is not the result of learning new data after having already learned an initial set of patterns, but rather the result of the network's learning a function 'too well', in some sense.

To understand the notion of catastrophic remembering, consider a network that learns to auto-associate a large number of patterns. The way in which the network 'knows' whether or not it has seen a particular pattern before is by comparing the pattern on input and on output – the result of its having passed through the network. If there is very little difference, it concludes that it already 'auto-associated' that particular pattern. In other words, it had already seen it. On the other hand, a large input–output difference means that it has encountered a new pattern. But now, consider what happens if the network has learned so many patterns that it has effectively learned the identity function. Once the network can reliably produce on output what it received on input for a large enough set of patterns, it will generalize correctly but it will 'remember' virtually any pattern, whether or not it has actually ever seen it before. The fundamental difficulty is that the network has then lost its ability to discriminate previously seen input from new input, even though it is generalizing the new input correctly. Thus, the ability to generalize to the identity function will necessarily mean that there will be a loss of discrimination.

The problem of catastrophic remembering remains an important one, and one for which current auto-associative connectionist memory models have no immediate answer.

Reference

a Sharkey, N. and Sharkey, A. (1995) An analysis of catastrophic interference *Connection Sci.* 7, 301–329

Representations in ALCOVE, depending on how finely the inverse-distance activation function is tuned, can vary from being somewhat distributed to highly local. When they are semi-distributed, this confers 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 does increases. In other words, if the receptive field of an input becomes restricted enough, the ALCOVE network becomes, for all intents and purposes, a localist network, thereby avoiding catastrophic interference from new input.

Distributed models that are sensitive and stable in the presence of new information

Certain models that rely on distributed, overlapping representations do not seem to forget catastrophically in the presence of new information. In particular, the class of convolution-correlation models, such as CHARM[34] and TODAM[35], and Sparse Distributed Memory (SDM)[36] can learn new information in a sequential manner and can, in addition, generalize on new input. The performance of these models on previously learned information declines gradually, rather than falling off abruptly, when learning new patterns. While, strictly speaking, convolution-correlation models and SDM are not 'connectionist' models, the former are readily shown to be isomorphic to sigma–pi connectionist models and SDM is isomorphic to a Hopfield network[37]. While there are critical storage limits to this type of memory (and therefore also in Hopfield networks) beyond which memory retrieval becomes abruptly impossible, below this limit SDM's internal representations precisely fit the bill of being semi-distributed. Their sparseness ensures a low degree of overlap, while their distributedness ensures that generalization will be maintained. In CHARM and TODAM, the input vectors comprise a large number of features that are bimodally coded, with an expected mean, over all features, of zero. This coding is critical and ensures a significant degree of orthogonality on input, which as we have seen[26], in general, decreases catastrophic forgetting.

Rehearsal of prior learning

Connectionist learning, especially in feedforward backpropagation networks, is a very contrived kind of learning. All of the patterns to be learned must be presented concurrently and repeatedly until the weights of the network gradually converge to an appropriate set of values. Real human learning, on the other hand, is largely sequential, even if it is true that many old items are refreshed continually in memory ('rehearsed') because we encounter them over and over. A number of authors[2,15,28,38] have studied various 'rehearsal' schemes to alleviate catastrophic interference. In this paradigm, learning is not truly sequential, rather, a number of the previously learned items are explicitly mixed in ('rehearsed') along with the new patterns to be learned. Numerous methods of choosing which of the previously learned items are to be mixed with the new patterns have been studied and, as expected, all were found to decrease the severity of catastrophic forgetting.

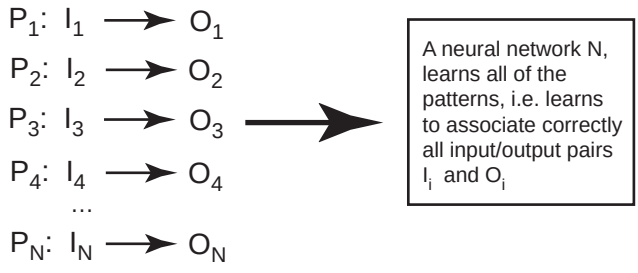
In 1995, Robins made a significant contribution to the field by introducing his 'pseudopattern' technique[38] (see Box 4) for doing rehearsal when none of the original patterns were available. This technique, combined with the notion of separate processing areas in the brain[39] led to the development of dual-network models discussed below[40,41].

Separating new learning from old learning

French[18,19] suggested that to alleviate catastrophic forgetting in distributed networks dynamical separation of their internal representations during learning was necessary. McClelland, McNaughton, and O'Reilly[39] went even further and suggested that nature's way of implementing this obligatory separation was the evolution of two separate areas of the brain, the hippocampus and the neocortex. They justified the brain's bimodal architecture by suggesting that the sequential acquisition of new data is incompatible with the gradual discovery of structure and can lead to catastrophic interference with what has previously been learned. In light of these observations, they suggested that the neocortex may be optimized for the gradual discovery of

Box 4. Approximating reality with 'pseudopatterns'

Mixing previously learned items ('rehearsal') with the new items to be learned has been shown to be an effective way to transform catastrophic forgetting into everyday gradual forgetting (Refs a–d). But what if the old patterns are unavailable for rehearsal? Robins developed a technique that significantly decreased catastrophic interference, even in cases where the previously learned patterns were not available for re-presentation to the network (Ref. d). Robins' idea was as simple as it was effective.



This set of inputs and outputs defines a function *f*

This network has now learned *f*

We now want the network to learn a set of New patterns, without forgetting the previously learned information. But we discover that the Original patterns are no longer available, and therefore cannot be interleaved with the New patterns to be learned.

- Question: Can we create substitute patterns to replace Original patterns?
- Answer: Yes, by generating a series of random input patterns $\hat{I}_i$, each of which is fed through the network to produce an output $\hat{O}_i$, as follows:

$$\psi_1: \hat{I}_1 \rightarrow \hat{O}_1$$
$$\psi_2: \hat{I}_2 \rightarrow \hat{O}_2$$
$$\psi_3: \hat{I}_3 \rightarrow \hat{O}_3$$
$$\psi_4: \hat{I}_4 \rightarrow \hat{O}_4$$
$$\ldots$$
$$\psi_M: \hat{I}_M \rightarrow \hat{O}_M$$

This set of pseudopatterns (see Ref. d) $\{\psi_1, \psi_2, \ldots, \psi_M\}$ approximates the originally learned function *f*.

These pseudopatterns will be interleaved with the New patterns and the network will learn the whole set of pseudopatterns along with the New patterns, thereby significantly reducing catastrophic forgetting of the originally learned information.

(Online: Fig. I)

After a network has learned a series of patterns, its weights encode a function, *f*, defined by those patterns. Now, if the original patterns are no longer available, how can we discover what *f* might have looked like, even approximately? Robins' solution was to 'bombard' the network inputs with random patterns ('pseudo-inputs') and observe the outputs generated by these random inputs. Each random input $\hat{\imath}$ that was fed through the network produced an output $\hat{o}$. The association $(\hat{\imath}, \hat{o})$ formed what Robins called a 'pseudopattern' (designated below by ψ) because, of course, the network had never previously actually seen the input $\hat{\imath}$. These pseudopatterns approximate the originally learned function, *f*, and can be interleaved with the new patterns to be learned to prevent catastrophic forgetting of the original patterns. So, just as rehearsing on previously learned patterns prevents a network from forgetting those patterns, rehearsing on pseudopatterns that approximate the function defined by the originally learned patterns also prevents catastrophic forgetting of the original patterns (although, of course, it doesn't work as well as rehearsing on the original patterns). Frean and Robins have developed some of the mathematics underlying the use of pseudopatterns to alleviate forgetting (Ref. e). The pseudopattern technique has also been successfully extended to Hopfield networks (Ref. f). Robins explored the possibility of pseudopatterns as a means by which memory consolidation occurs (Ref. g). This technique has also been used successfully as the means of information transfer between storage areas in dual-network memory models (Refs h,i).

One important question in this area is how best to optimize the pseudopatterns used to recover information. Are there ways to improve 'quality' of the pseudopatterns so that they better reflect the originally learned regularities in the environment?

References

a 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,* pp. 26–33, Erlbaum

b Ratcliff, R. (1990) Connectionist models of recognition memory: constraints imposed by learning and forgetting functions *Psychol. Rev.* 97, 285–308

c Murre, J. (1992) *Learning and Categorization in Modular Neural Networks,* Erlbaum

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

e Frean, M. and Robins, A. (1998) Catastrophic forgetting and 'pseudorehearsal' in linear networks, in *Proceedings of the Ninth Australian Conference on Neural Networks* (Downs, T., Frean, M. and Gallagher, M., eds), pp. 173–178, University of Queensland

f Robins, A. and McCallum, S. (1998) Pseudorehearsal and the catastrophic forgetting solution in Hopfield-type networks *Connection Sci.* 10, 121–135

g Robins, A. (1996) Consolidation in neural networks and in the sleeping brain *Connection Sci.* 8, 259–275

h French, R.M. (1997) Pseudo-recurrent connectionist networks: an approach to the 'sensitivity–stability' dilemma *Connection Sci.* 9, 353–379

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

the shared structure of events and experiences, and that the hippocampal system is there to provide a mechanism for rapid acquisition of new information without interference with previously discovered regularities. After this initial acquisition, the hippocampal system serves as a teacher to the neocortex.

The earliest attempt at implementing a dual-net architecture in order to decrease catastrophic interference was a model called Jumpnet[42]. This model consisted of a standard connectionist network (the 'processing' network) coupled with a 'control' network that modulated the weights of the processing network. While this type of dual network was shown to reduce catastrophic interference in certain cases, it was not clear that it could effectively handle the type of problem most likely to cause catastrophic forgetting – namely, learning new patterns whose inputs are very similar to that of previously learned patterns, but whose outputs are quite different.

French[40] and Ans and Rousset[41] independently developed dual-network architectures based on the principle of two separate pattern-processing areas, one for early-processing, the other for long-term storage. In both models, the

early-processing and storage areas are in continual communication, transferring information back and forth by means of pseudopatterns[38]. Both models exhibit gradual forgetting and, consequently, plausible sequence learning. In French's pseudo-recurrent network[40], the pseudopattern transfer mechanism leads to a gradual 'compression' (i.e. fewer active nodes) of internal representations in the long-term storage area. It has been shown that the representational compression inherent in this kind of dual-network system, designed to reduce catastrophic interference, would produce certain patterns of category-specific deficits actually observed in amnesiacs[43]. It has also been shown that in human list-learning, adding new items to the list decreases recall of earlier items in the list[44,45]. By contrast, strengthening of particular items (for example, by repeating them) does not produce decreased recall of the unstrengthened items (i.e. there is no so-called list-strength effect)[46]. The pseudo-recurrent architecture, like humans, exhibits a plausible list-length effect and the absence of a list-strength effect, a dissociation that causes problems for many current connectionist models.

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, Chappell and Humphreys[47] combined an auto-associative architecture with sparse representations to successfully reduce the level of catastrophic interference. Like the dual-network architectures, their architecture also exhibits a list-length and no list-strength effect. Hinton and Plaut[12] were able to reduce interference in new learning by using two different kinds of weights instead of one. One set changed rapidly, but decayed to zero rapidly ('fast' weights); the other was hard to change, but decayed only slowly back to zero ('slow' weights). The weight used in the learning algorithm was a combination of slow and fast weights. This technique, although frequently cited, has not yet been thoroughly explored, although it is likely that there are storage capacity limitations to this type of solution. In other words, while it can be used to mitigate the influence of one or two new patterns on previously learned patterns, is the technique sufficiently powerful to permit true sequential learning similar to that in dual-network architectures? Another more recent dual-weight architecture[48] employs two sets of independent weights and taxonomically falls somewhere between dual-network models[40,41] and single-network, dual-weight architectures[12]. Cascade-correlation[49] has also been tried as a means of alleviating catastrophic interference with some success[50].

Outstanding questions

- Do all models that exhibit gradual forgetting rather than catastrophic forgetting necessarily rely on some form of representational separation?
- Are dual-network systems really necessary for the brain to overcome the problem of catastrophic forgetting?
- How does episodic memory fit into this picture?
- Does the pseudopattern mechanism proposed by Robins really have a neural correlate? If so, are neural pseudopatterns produced, say, during REM sleep? And are they really as random as the pseudopatterns used in present dual-network connectionist models or has the brain evolved a better way of doing 'rehearsal' in the absence of real input from the environment?
- How close can we get to the ideal of good generalization, good discrimination, immunity to catastrophic interference and good episodic memory, in a single, distributed system?
- What types of animals are subject to catastrophic interference and under what circumstances? Are there circumstances under which humans do experience catastrophic forgetting?

Conclusion

For nearly a decade researchers have been studying the problem of catastrophic interference in connectionist networks. Modeling true sequential learning of the kind that we humans do requires appropriate solutions of this problem to be found. Recent research seems to indicate that one possible solution to the problem is two separate, permanently interacting processing areas, one for new information, the other for long-term storage of previously learned information. Even though it is far from obvious that this is the only way to handle the problem of catastrophic forgetting, it has been argued that this is how the human brain evolved to deal with the problem. Further research may reveal whether this is, in fact, the case.

Acknowledgements

Thanks to André Ferrara for the research reported in Box 1. The present paper was supported in part by research grant IUAP P4/19 from the Belgian government.

References

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

2 Ratcliff, R. (1990) Connectionist models of recognition memory: constraints imposed by learning and forgetting functions *Psychol. Rev.* 97, 285–308

3 Minsky, M. and Papert, S. (1969) *Perceptrons*, MIT Press

4 Rosenblatt, F. (1958) The perceptron: a probabilistic model for information storage and organization in the brain *Psychol. Rev.* 65, 386–408

5 Rosenblatt, F. (1962) *Principles of Neurodynamics*, Spartan, NY, USA

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

7 Carpenter, G. and Grossberg, S. (1987) ART 2: self-organization of stable category recognition codes for analog input patterns *Appl. Opt.* 23, 4919–4930

8 Barnes, J. and Underwood, B. (1959) 'Fate' of first-learned associations in transfer theory *J. Exp. Psychol.* 58, 97–105

9 Rumelhart, D., Hinton, G. and Williams, R. (1986) Learning representations by back-propagating error *Nature* 323, 533–536

10 Carpenter G. and Grossberg, S. (1986) Adaptive resonance theory: stable self-organization of neural recognition codes in response to arbitrary list of input patterns, in *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pp. 45–62, Erlbaum

11 Hinton, G., McClelland, J. and Rumelhart, D. (1986) Distributed representations, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Vol. 1. Foundations* (Rumelhart, D. and McClelland, J., eds), pp. 77–109, MIT Press

12 Hinton G. and Plaut D. (1987) Using fast weights to deblur old memories, in *Proceedings of the Ninth Annual Conference of the Cognitive Science Society,* pp. 177–186, Erlbaum

13 Sutton, R. (1986) Two problems with backpropagation and other steepest-descent learning procedures for networks, in *Proceedings of the Eighth Annual Conference of the Cognitive Society,* pp. 823–831, Erlbaum

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

15 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*, pp. 26–33, Erlbaum

16 Ebbinghaus, H. (1885) Über das Gedächtnis: Untersuchen zur Experimentellen Psychologie ('On memory') (H.A. Ruger and C.E. Bussenius, transl. 1964), Dover

17 Kortge, C. (1990) Episodic memory in connectionist networks, in *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, pp. 764–771, Erlbaum

18 French, R.M. (1991) Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks, in *Proceedings of the 13th Annual Cognitive Science Society Conference*, pp. 173–178, Erlbaum

19 French, R.M. (1992) Semi-distributed representations and catastrophic forgetting in connectionist networks *Connection Sci.* 4, 365–377

20 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,* pp. 380–387, Erlbaum

21 McRae, K. and Hetherington, P. (1993) Catastrophic interference is eliminated in pretrained networks, in *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, pp. 723–728, Erlbaum

23 Sharkey, N. and Sharkey, A. (1995) An analysis of catastrophic interference *Connection Sci.* 7, 301–329

24 French, R.M. (1994) Dynamically constraining connectionist networks to produce distributed, orthogonal representations to reduce catastrophic interference, in *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, pp. 335–340, Erlbaum

25 Lewandowsky, S. and Shu-Chen Li (1993) Catastrophic interference in neural networks: causes, solutions, and data, in *New Perspectives on Interference and Inhibition in Cognition* (Dempster, F.N. and Brainerd, C., eds), pp. 329–361, Academic Press

26 Lewandowsky S. (1991) Gradual unlearning and catastrophic interference: a comparison of distributed architectures, in *Relating Theory and Data: Essays on Human Memory in Honor of Bennet B. Murdock* (Hockley, W. and Lewandowsky, S., eds), pp. 445–476, Erlbaum

27 Murre, J. (1992) The effects of pattern presentation on interference in backpropagation networks, in *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, pp. 54–59, Erlbaum

28 Murre, J. (1992) *Learning and Categorization in Modular Neural Networks*, Erlbaum

29 Krushke, J. (1992) ALCOVE: an exemplar-based model of category learning *Psychol. Rev.* 99, 22–44

30 Kruschke, J. (1993) Human category learning: implications for backpropagation models *Connection Sci.* 5, 3–36

31 Sloman, S. and Rumelhart, D. (1992) 'Reducing interference in distributed memories through episodic gating', in *Essays in Honor of W. K. Estes* (Healy, A., Kosslyn, S. and Shiffrin, R., eds), pp. 227–248, Erlbaum

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

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

34 Metcalfe, J. (1982) A composite holographic associative recall model *Psychol. Rev.* 89, 627–661

35 Murdock, B. (1983) A distributed memory model for serial-order information *Psychol. Rev.* 100, 183–203

36 Kanerva, P. (1989) *Sparse Distributed Memory*, MIT Press

37 Keeler, J.D. (1988) Comparison between Kanerva's SDM and Hopfield-type neural networks *Cognit. Sci.* 12, 279–329

38 Robins, A. (1995) Catastrophic forgetting, rehearsal, and pseudo-rehearsal *Connection Sci.* 7, 123–146

39 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 *Psychol. Rev.* 102, 419–457

40 French, R.M. (1997) Pseudo-recurrent connectionist networks: an approach to the 'sensitivity–stability' dilemma *Connection Sci*. 9, 353–379

41 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

42 Rueckl, J. (1993) Jumpnet: a multiple-memory connectionist architecture, in *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, pp. 866–871, Erlbaum

43 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 20th Annual Cognitive Science Society Conference*, pp. 374–379, Erlbaum

44 Ratcliff, R. and Murdock, B. (1976) Retrieval processes in recognition memory *Psychol. Rev.* 83, 190–214

45 Strong, E. (1912) The effect of length of series upon recognition memory *Psychol. Rev.* 19, 447–462

46 Murnane, K. and Shiffrin (1991) Interference and the representation of events in memory *J. Exp. Psychol: Learn. Mem. Cognit.* 17, 355–374

47 Chappell M. and Humphreys, M. (1994) An auto-associative neural network for sparse representations: analysis and application to models of recognition and cued recall *Psychol. Rev.* 101, 103–128

48 Levy J. and Bairaktaris, D. (1995) Connectionist dual-weight architectures *Lang. Cognit. Process.* 10, 265–283

49 Fahlman, S. and Libiere, C. (1990) The cascade-correlation learning architecture, in *Advances in Neural Information Processing Systems* (Vol. 2) (Touretsky, D., ed.), pp. 524–532, Morgan-Kaufmann

50 Tetewsky, S.J., Shultz, T.R. and Takane, Y. (1995) Training regimens and function compatibility: implications for understanding the effects of knowledge on concept learning, in *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pp. 304–309, Erlbaum