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 Cohen1 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 Papert3 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’ problem4. 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 remaining 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 Underwood8 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 Cohen1 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 network9 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...
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 net-
work 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 ‘catastrophi-
cally’ forgotten its one’s addition sums. In a subsequent ex-
pertiment that attempted to more closely match the original
Blames and Underwood paradigm, they again found the
same catastrophic, rather than gradual, forgetting in the
neutral network they tested. Ratcliff1 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 cata-
strophically 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 Cohen1 answered this as fol-
lows: ‘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:
how 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 ani-
mals with a less highly evolved brain organization? (see Box 1).

And finally, will distributed connectionist networks
remain unable to perform true sequential learning?2 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, forget-
ing of old, unrepresented patterns occurs gradually over time.

However, for any network subject to catastrophic inter-
fERENCE, learning cannot occur in this manner, because the
new learning will effectively erase previous learning.

Measuring catastrophic interference

The two initial studies on catastrophic interference1,2 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 recog-
nition 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 pat-
tern. Hetherington and Siedenberg13 introduced a ‘savings’
measure of forgetting based on a relearning measure first
proposed by Ebbinghaus14. 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. Kortge2 claimed that
the problem was not one inherent in distributed connec-
tionist architectures, but rather was due to the backpropa-
gation learning rule. He developed a variation of the back-
propagation algorithm using what he called ‘novelty vec-
tors’ that produced a decrease in catastrophic interfer-
ence. The idea is that ‘when the network makes an error, we
would like to blame just those active units which were ‘re-
 sponsible’ for the error – blaming any others leads to excess
interference with other patterns’ output’. When a new pat-
tern 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 in-
tended 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 cor-
responding 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.
French2,5,7 suggested that, in general, catastrophic forget-
ting was largely a consequence of the overlap of internal dis-
tributed 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 repre-
sentational 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 acti-
vation sharpening algorithm6,8. An extra step is added to the
standard backpropagation learning algorithm in which activa-
tions patterns at the hidden layer are ‘sharpened’, that is, the
activation level of the most active hidden node(n) is
increased slightly for each pattern, while the activations of
other nodes are decreased. This technique had the effect of
’sparifying’ 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 periodic 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 time of lever-pressing 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 evidence of a step from the initial 40-second learning. One reasonable interpretation of this result is that the 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 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.

Reducing representational overlap

In one way or another, almost all of the early techniques relied on reducing representational overlap. Some attempted to use orthogonal encoding of inputs26,27. These techniques used bipolar feature coding (i.e., −1) for 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 patterns28,29. It turned out that internal orthogonalization of representations could be made to emerge automatically by pre-training30. 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 human and...
When a distributed network has learned to recognize an initial set of patterns, this means that it has found a point in weight-space, $W_{\text{old}}$, 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, $W_{\text{new}}$, 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 already 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 small problem (SOR) (see Fig. 1). It is immediately obvious why this would imply that a move from $W_{\text{old}}$ to $W_{\text{new}}$ could have catastrophic effects on the previously learned knowledge, both in terms of the exact-recognition and retraining-time criteria for forgetting.

References

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

When a distributed network has learned to recognize an initial set of patterns, this means that it has found a point in weight-space, $W_{\text{old}}$, 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, $W_{\text{new}}$, 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 already 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 small problem (SOR) (see Fig. 1). It is immediately obvious why this would imply that a move from $W_{\text{old}}$ to $W_{\text{new}}$ could have catastrophic effects on the previously learned knowledge, both in terms of the exact-recognition and retraining-time criteria for forgetting.

References


c Kolen, J. and Pollack, J. (1990) Backpropagation is sensitive to initial conditions Complex Syst. 4, 269–288

d Rumelhart, D. (1988) ‘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.)

Fig. 1. 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 cliff’ that represent the unpredictable outcome for convergence of even small movements in weight space for such networks. (Reprinted, with permission, from Ref. c.)
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 generating 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. 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 authors37–40 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 this field by introducing his ‘pseudopattern’ technique41 (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 brain42 led to the development of dual-network models discussed below43,44.

Separating new learning from old learning

French18,19 suggested that to alleviate catastrophic forgetting in distributed networks dynamical separation of their internal representations during learning was necessary. McClelland, McNaughton, and O’Reilly44 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 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 increase. 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 cataclysmically in the presence of new information. In particular, the class of convolution–correlation models, such as CHARMM and TODAM43, and Sparse Distributed Memory (SDM)44 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 network45. While there are critical storage limits to this type of memory (and therefore also to 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, in general, decreases catastrophic forgetting.

Reference


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 generating 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.
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-learned regularities. After this initial acquisition, the hippocampal system serves as a teacher to 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 and Ans 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

### Box 4. Approximating reality with ‘pseudopatterns’

Mixing previously learned items (‘rehearsed’) 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.

![A neural network N. learns all of the patterns, i.e. learns to associate correctly all input/output pairs I, and O.](image)

After a network has learned a series of patterns, its weights encode a function, ƒ, defined by those patterns. Now, if the original patterns are no longer available, how can we discover what ƒ 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 ɨ was fed through the network produced an output ŷ. The association (ɨ, ŷ) formed what Robins called a ‘pseudopattern’ (designated below by ψ). Then, of course, the network had never previously actually seen the input ɨ. These pseudopatterns approximate the originally-learned function, ƒ, 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). French 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). French 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 optimise the pseudopatterns used to recover information. Are these two ways to improve ‘quality’ of the pseudopatterns so that they better reflect the originally-learned regularities in the environment?
early-processing and storage areas are in continual communication, transferring information back and forth by means of pseudopatterns. Both models exhibit gradual forgetting and, consequently, plausible sequence learning. In French’s pseudo-sequential network, 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. It has also been shown that in human list-learning, adding new items to the list decreases recall of earlier items (the list-strength effect). 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). The pseudo-sequential 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 Humphries2 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 Plaut3 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, the technique sufficiently powerful to permit true sequential learning similar to that in dual-network architectures? Another more recent dual-weight architecture4 employs two sets of independent weights and taxonomically falls somewhere between dual-network models5,6 and single-network, dual-weight architectures. Cascade-correlation7 has also been tried as a means of alleviating catastrophic interference with some success8.

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.

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?

References


Acknowledgements

Thanks to André Ferrara for the research reported in Box 1. The present paper was supported in part by research grant UAP/F19 from the Belgian government.

Trends in Cognitive Sciences – Vol. 3, No. 4, April 1999

FRENCH – Catastrophic forgetting


Hetherington, P. (1991) The Sequential Learning Problem, Master’s Thesis, Department of Psychology, McGill University, Montreal, Quebec, Canada
Marcus, J. (1992) Learning and Computation in Modular Neural Networks, Erlbaum
McNaughton, B. (1988) A distributed memory model for serial-order information, Psychol. Rev. 95, 98–107

Coming soon to Trends in Cognitive Sciences

- Does generalization in infant learning implicate abstract algebra-like rules? by J.L. McClelland and D.C. Plaut
- Response from G. Marcus
- Is imitation learning the way to humanoid robots? by S. Schaal
- Models of word production, by W.J.M. Levelt
- Imaging visual recognition: PET and fMRI studies of the functional anatomy of human visual recognition, by M.J. Farah and G.K. Aguirre

Review
French – Catastrophic forgetting

Trends in Cognitive Sciences – Vol. 3, No. 4, April 1999