Solutions to the Catastrophic Forgetting Problem

Anthony Robins (COSCAVR@OTAGO.AC.NZ)
Computer Science Department, The University of Otago,
P.O. Box 56, Dunedin, New Zealand

Abstract

In this paper we review three kinds of proposed solutions to the catastrophic forgetting problem in neural networks. The solutions are based on reducing hidden unit overlap, rehearsal, and pseudorehearsal mechanisms. We compare the methods and identify some underlying similarities. We then briefly note some potential implications of the rehearsal / pseudorehearsal based methods, including their application to sequential learning tasks.

Introduction

In most standard neural network learning algorithms, such as back-propagation (Rumelhart, Hinton & Williams, 1986), all information is learned "concurrently". In other words, the whole population of interest is presented and trained as a single, complete entity. Training is then "finished" and no further information is added to the network. Being limited to concurrent learning is undesirable in practical terms, making it very difficult to modify or extend any given neural network application without completely retraining the network (compared with a traditional rule based system where information or rules can easily be added to or removed from the system). It is also a highly implausible constraint for cognitive modelling where so much of human learning is clearly sequential or incremental in nature. This limitation arises because of the "catastrophic forgetting" problem – the learning of new information disrupts previously learned information in a network.

In this paper we review three kinds of proposed solutions to the catastrophic forgetting problem. They are based on reducing hidden unit overlap, rehearsal, and pseudorehearsal mechanisms. We compare the methods and identify some underlying similarities. Rehearsal and pseudorehearsal allow new information to be added to a network sequentially (at any time) without disrupting old information. We briefly explore some potential implications of these methods, including the possibility of a framework for modelling ongoing or continuous learning / development with neural networks, and speculations about the relationship of these methods to the consolidation of information during sleep.

Catastrophic forgetting and concurrent learning

Ideally the representations developed by a learning system should be stable enough to preserve important information over time, but plastic enough to incorporate new information when necessary. One consequence of a failure to address this "stability / plasticity dilemma" (Grossberg, 1987) in many neural networks is excessive plasticity, usually called "catastrophic forgetting" (or "catastrophic interference", or the "serial learning problem"). If a network is exposed to the learning of new information, then any previously learned information will typically be greatly disrupted or lost. Grossberg (1987) suggests the analogy of a human trained to recognise the word "cat", and subsequently to recognise "table", being then unable to recognise "cat".

A number of recent studies have used multi-layer perceptron (typically back-propagation) networks to highlight the problem of catastrophic forgetting and explore various issues – these include, McCloskey & Cohen (1989), Hetherington & Seidenberg (1989), Ratcliff (1990), Lewandowsky (1991), Murre (1992a, 1992b), French (1992, 1994, 1997), McRae & Hetherington (1993), Lewandowsky & Li (1995), Sharkey & Sharkey (1995), Robins (1995, 1996a), and Frean & Robins (1997). Similar issues have been explored in the context of Hopfield networks by Nadal, Toulouse, Changeux & Dehaene (1986), Burgess, Shapiro & Moore (1991), and Robins & McCallum (1998).

In a typical illustration of catastrophic forgetting we use a back-propagation network to learn a base population of items (input / output vector pairs) in the usual way. Subsequently a number of new items are learned one by one[1]. The effect of these new items can be illustrated by plotting a measure (such as goodness or error) of the ability of the network to correctly reproduce the base population after each new item. As shown in Figure 1, the error in a base population of items increases "catastrophically" after the learning of even one new item, and continues to rise as further new items are learned.

This catastrophic forgetting is the underlying constraint that restricts most neural networks to concurrent learning (where the whole population of interest must be learned as a single, complete entity).

[1] All simulations in this paper use the "Iris" data set (Murphy & Aha, 1994) consisting of 150 items divided into three classes (distinct species of iris) of 50 items each. Each item consists of four real valued measurements of the iris (such as petal length). We used a 4:3:4 or 4:4:4 autoassociative back-propagation network with a learning constant of 0.05 and a momentum constant of 0.9, and an error criterion of 0.01. All results reported were averaged over 50 individual replications of the simulation (using different populations for each replication).

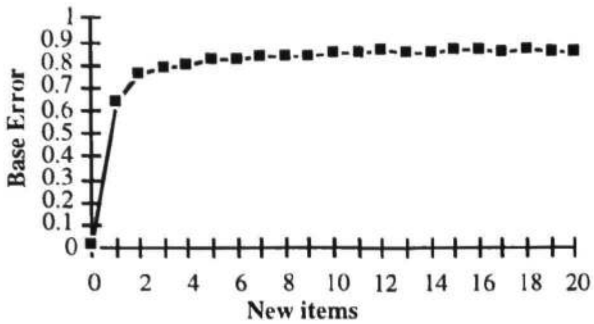

Figure 1: The basic catastrophic forgetting effect. (Adapted from Robins (1996) Figure 1).

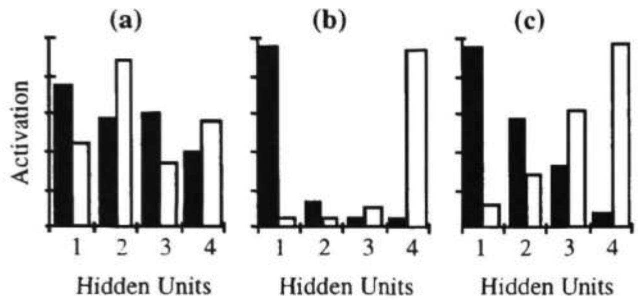

Figure 2: Hypothetical hidden unit activations.

Solutions to catastrophic forgetting

In this section we briefly review three general approaches to solving the catastrophic forgetting problem. These are based on reducing hidden unit overlap, on rehearsal, and on pseudorehearsal respectively. Other proposed solutions based on specific "purpose built" architectures are noted in Sharkey and Sharkey (1995).

Reducing hidden unit overlap

French (1992) suggests that the extent to which catastrophic forgetting occurs is largely a consequence of the overlap of distributed representations, and that the effect can be reduced by reducing this overlap. Catastrophic forgetting will be worst when new item inputs are similar to base population inputs (i.e. generate similar hidden unit patterns) but require very different output patterns to be produced.

Several studies have explored mechanisms for reducing representational overlap and their impact on catastrophic forgetting. The novelty rule (Kortge, 1990), activation sharpening (French, 1992), and techniques developed by Murre (1992a) and McRae and Hetherington (1993) all fall within this general framework. These methods focus on increasing the separation (orthogonality) of the hidden unit representations developed by the network, typically by creating "sparser" representations (hidden unit patterns with a smaller number of active units). French's activation sharpening, for example, introduces an extra step to the learning process for each input which adds weight changes that slightly increases the activation of the most active (or k most active) hidden units while decreasing the activations of all others. To illustrate these points consider a hypothetical network with four hidden units which has learned a

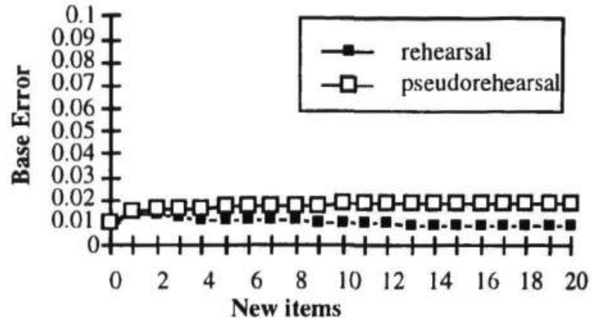

Figure 3: Rehearsal and pseudorehearsal methods effectively eliminate the catastrophic forgetting. (Adapted from Robins (1996) Figures 3 & 4).

population where the inputs are divided into two relatively distinct categories. Figure 2 (a) illustrates possible hidden unit patterns of activation for each category in a standard back-propagation network. The representations of each category may not be well separated. Figure 2 (b) illustrates typical hidden unit patterns of activation for each category in a back-propagation network using single node activation sharpening. The representations of each category are well separated, and sparse.

While these methods can reduce the impact of catastrophic forgetting in varying ways (as discussed below), French (1994) identifies several problems with the use of sparse hidden unit representations. French shows that the reduced representational capacity of sparse patterns of activation can in some circumstances result in an *increase* in catastrophic forgetting, and also argues that it results in a reduced capacity to categorise and discriminate inputs, and a reduced capacity to generalise. The use of sparse (more "localist") representations would also imply a decreased robustness in the face of noise and damage. French (1994) concludes that hidden layer representations need to be highly distributed as well as separated, and describes a new method, "context biasing", which generates such representations. Context biasing introduces an extra step to the learning process for each input which adds weight changes that enhance the differences between the pattern of hidden unit activation generated by the current input and the pattern generated by the previous input. Figure 2 (c) illustrates typical hidden unit patterns of activation for each of two hypothetical categories in a back-propagation network using context biasing. The representations of each category are well separated, but distributed, avoiding the problems associated with sparse representations.

Of the methods for reducing hidden unit overlap, activation sharpening and context biasing do not actually prevent a base population from being disrupted by new items. These methods do, however, ameliorate the effects of this catastrophic forgetting to the extent that they allow the base population to be subsequently retrained to criterion more quickly than is the case in a standard back-propagation network. The novelty rule (Kortge, 1990) has been shown to prevent catastrophic forgetting, but can only be used with autoencoder (autoassociative) networks. Catastrophic forgetting may also be prevented if it is possible to pre-train the network on a population which is "relevant" to the base population and new items (simulating prior knowledge of a task domain) as explored by Sharkey and Sharkey (1995) and McRae and Hetherington (1993). In McRae and Hetherington's simulations this pretraining naturally reduced the overlap of hidden unit representations in subsequent learning.

Rehearsal

A second general approach to preventing catastrophic forgetting involves "rehearsing" the base population by retraining some base population items as the new items are trained. Ideally this will allow the new items to be incorporated into the structure of the base population instead of just overwriting it. Rehearsal was first explored in the context of catastrophic forgetting by Hetherington and Seidenberg (1989) and Ratcliffe (1990), and a range of rehearsal methods have been explored by Murre (1992b) and Robins (1995).

Following Ratcliffe (1990), rehearsal can be thought of as introducing each new item not alone, but in a *rehearsal buffer* along with a number of old items. The population of items in the rehearsal buffer are then trained over a number of epochs (iterations of the learning algorithm) in the usual way. The various possible ways of selecting and managing the old items in a rehearsal buffer define a family of possible *rehearsal regimes*. Robins (1995) explores a range of rehearsal regimes, including a "recency" regime (following Ratcliffe (1990)), a "random" regime (independently proposed in Murre (1992b)), and a "sweep" regime. In this paper we will illustrate the general properties of rehearsal using the sweep regime as a specific example. In sweep rehearsal the rehearsal buffer always contains the new item, and also contains a number (one or more) of old items that are randomly selected *for each epoch* of training (replacing the old items used in the previous epoch so that the buffer remains of a fixed size)[2]. Training continues until the single new item reaches criterion.

Our second simulation explores the performance of (sweep) rehearsal compared to the simple no rehearsal condition illustrated in the first simulation (see Figure 1). We use the same network, parameters, and populations, i.e. a base population consisting of 30 items of one species of iris and 20 new items drawn from a second species. Each new item is trained in a buffer along with a number (five in this case) of previously learned items (base population items or new items learned earlier in the sequence) chosen at random for each epoch. The results are shown in Figure 3, "Rehearsal" condition. Performance on the base population is maintained very effectively.

In rehearsal one chooses some number of old items to be learned alongside a new item. If *all* old items were included, rehearsal would simply amount to retraining the entire base population as new items are introduced (as is the case in for example the "interleaved learning" proposed by McClelland, McNaughton and O'Reilly (1995)). What is interesting about the studies described above, however, is that subsets of the base population or less rigorous training criteria can also be used effectively. The (sweep) rehearsal regime illustrated here is very effective despite the fact that it does not use all the items at every step and does not explicitly retrain old items to criterion. This suggests that in general rehearsal should be "broad" but it does not need to be "deep".

Pseudorehearsal

Rehearsal of this kind is an effective solution as long as the previously learned items are actually available for relearning. It may be, however, that the old items have been lost, or it is not practical for some reason to store them. Sharkey and Sharkey note, for example, that:

> "the interference [catastrophic forgetting] problem is [...] general and should be of concern to all those involved in developing applications in which the training data only become available piecemeal over an extended period of time. For example, in on-line learning of control processes, such as found in robotics or manufacturing, it may not be practical to maintain all of the training data in memory and retrain each time a novel aspect of the data is encountered." Sharkey & Sharkey (1995, p 302).

In any case, retaining old items for rehearsal in memory seems somewhat artificial, as it requires that they be available on demand from some other source, which would seem to make the memory itself redundant!

It is possible to achieve the benefits of rehearsal, however, even when there is no access to the base population. In other words, we can do rehearsal even when we do not have the old items to rehearse! This "pseudorehearsal" mechanism, introduced in Robins (1995), is based on the use of artificially constructed populations of "pseudoitems" instead of the actual old items.

A pseudoitem is constructed by generating a new input vector at random, and passing it forward through a network in the standard way. Whatever output vector this input generates becomes the associated target output. For a given network (trained on the base population) a population of pseudoitems constructed in this way can be used instead of the actual base population items in any rehearsal regime. Such a population is constructed before each new item is learned[3]. Learning proceeds exactly as before, except that instead of rehearsing items chosen from the old base population they are chosen from the population of pseudoitems.

Just as for simple rehearsal, in this paper we use the "sweep" regime for choosing pseudoitems (see Robins (1995) for other variants). Using the same network and

[2] The ratio of old items in the buffer to the size of the base population is an important factor. Simulations based on the Iris population in this paper continue to use the baseline established in Robins (1995) of setting the size of the rehearsal buffer to include a number of old items equal to roughly 15% of the size of the base population. This figure appears to provide an acceptable tradeoff between performance and the amount of rehearsal required. The performance of all regimes can be arbitrarily improved by increasing the size of the rehearsal buffer.

[3] Our simulations use populations of 128 pseudoitems.

populations as above, we repeat the training procedure of the rehearsal process (second simulation) except that pseudoitems are used instead of actual old items. Specifically, before each new item is learned a population of pseudoitems is constructed. The new item is then learned alongside pseudoitems (five in this case) that are chosen at random for each epoch. Training continues in this way until the new item is trained to criterion. The results are shown in Figure 3, "Pseudorehearsal" condition. Pseudorehearsal remains highly effective at preserving performance on the base population. After the twentieth new item, the error is roughly two percent of the error of the no rehearsal condition (Figure 1) and increasing only gradually.

In short, pseudorehearsal is a promising method for achieving the benefits of rehearsal in reducing catastrophic forgetting without assuming access to old information. Rather than explicitly storing all learned items for later rehearsal, pseudorehearsal *approximates* this information whenever it is needed. As well as autoassociative learning with the Iris data set used in this paper and Robins (1996), pseudorehearsal based mechanisms have been shown to be effective on: autoassociative and heteroassociative randomly constructed data sets by Robins (1995) and Ans & Rousset (1997); a classification task using the Mushroom data set (see Murphy & Aha (1992)) by French (1997); and an autoassociative alphanumeric character set (using a Hopfield type network) by Robins & McCallum (1998).

Pseudorehearsal is based on sampling the function fit by the network to the base population in the process of learning. Obviously the performance of pseudorehearsal based methods will be greatly influenced by the nature of this learning process. Networks which have been trained so as to generalise well (fit the base population data points with a smooth, compact function) will in general generate useful pseudoitems that preserve the structure of the base population well. Networks which do not generalise well (fit the base population data points with a noisy function) will not necessarily generate useful pseudoitems. As good generalisation is frequently a specific objective of training, however, there are a wide range of techniques which can be applied to constrain a network to learn compact functions (see for example Moody (1994)).

Comparing the methods

The essence of preventing catastrophic forgetting is to *localise changes* to the function learned by the network. Rehearsal accomplishes this by relearning the original training data points during new training. Pseudorehearsal accomplishes this by relearning other points randomly chosen from the function during new training (see Frean & Robins (1997) for further discussion). In short, rehearsal / pseudorehearsal works *directly* with the function to localise changes. Methods based on reducing the overlap of hidden unit patterns work *indirectly* by manipulating the "representation" of the function within the network. The two approaches are related in that some sharpening of hidden unit representations emerges naturally from the rehearsal / pseudorehearsal process.

In order to explore hidden unit representations this simulation uses the Iris population and a back-propagation network of the same architecture and parameters as the simulations above. Once again the network is trained on a

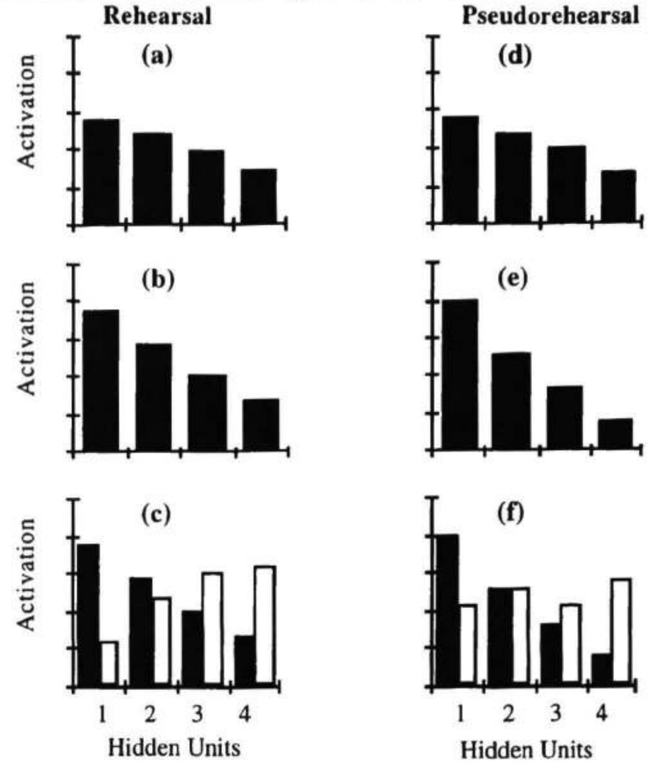

Figure 4: Hidden unit activations.

base population of 30 irises drawn from one species, and subsequently on 20 individual irises drawn from a second species.

The hidden unit patterns of activation after the learning of the base population and after subsequent learning of the new items are represented in Figure 4. Each graph shows the representations of the base population (filled bars) after learning, where a representation consists of the activation of the network's 4 hidden units arranged in order of most to least active (averaged over all items in the population). For the rehearsal condition (graphs (a) to (c)), graph (a) shows the representation of the base population after its initial learning. Graph (b) shows that this representation has been somewhat sharpened after the 20 new items have been added to the network using rehearsal. Graph (c) shows the same representation of the base population as graph (b) and contrasts it with the representation of the 20 learned new items (unfilled bars). Graphs (d) to (f) show equivalent results using pseudorehearsal instead of rehearsal.

For both rehearsal and pseudorehearsal conditions the subsequent learning of the new item population results in a somewhat "sharper" representation of the base population. Particularly in the rehearsal condition, the representations of both the base and new item populations (see Figure 4 (c) and (f)) have the same form as hidden unit representations generated by French's (1994) context biasing method (see Figure 2(c)), being usefully "distributed but separated". We suggest that in general the tendency to develop distributed but separated hidden unit representations will emerge naturally from rehearsal based processes. This "localisation of representation" may be one of the mechanisms by which the rehearsal processes achieve local changes to the base population function in a neural network.

A reduction in overlap also emerges naturally from McRae and Hetherington's (1993) pretraining method. This may account for the fact that both pretraining and rehearsal are able to actually prevent catastrophic forgetting, whereas in general techniques that directly modify the learning algorithm just ameliorate its effects as noted above. A relevant observation from our current simulation is that in only a minority of cases (11 out of 50 replications of the simulation for rehearsal, and 13 out of 50 replications for pseudorehearsal) are the two hidden units that are most active after the base population has been learned (i.e. parts (a) and (d) in Figure 4) the same units as the two most active units after the new items have also been learned (i.e. parts (b) and (e) in Figure 4). In general, then, the representation developed during the rehearsal process has involved a significant *re-ordering* of the units as well as an overall sharpening. This suggests that considerable flexibility may be needed to develop appropriate hidden unit representations, whereas any modification to the learning algorithm that directly sharpens patterns of activation works against flexibility by further entrenching established patterns.

French (1994) notes that very sparse representations may generalise poorly, and this is one of the motivations for his use of context biasing to develop distributed but separated representations. The similar representations emergent from the rehearsal and pseudorehearsal processes not only preserve the base population, but they also maintain good generalisation performance (as is characteristic of both neural networks and human cognition) on that base population. During the training and testing of the base population described above the performance of the network on a test population was also assessed. The test population consisted of a further 20 examples drawn from the same species of iris as the 30 base population items. Every time the average error of the base population was computed (i.e. for each trial from 0 to 20 new items) the average error of the test population was also computed. For both rehearsal and pseudorehearsal the error of the test population over all 21 trials typically exceeded the average error of the base population by no more than 0.005. Note that it is not the case that the networks are generalising well simply because they have learned to autoassociate any input (see the discussion of discriminability in Robins (1995) and Sharkey and Sharkey (1995)). Networks trained on all 150 irises to criterion and subsequently tested on 150 randomly constructed autoassociative items produce an average error of 0.158 for the random population (cf. 0.01 for the Iris population).

Discussion

To summarise, catastrophic forgetting is a natural consequence of an neural network style of learning and affects a wide range of networks. One family of solutions has been proposed which focuses on reducing the overlap of hidden unit representations. Some of these methods are effective at reducing catastrophic forgetting in specific circumstances, others reduce it in the sense that the disrupted base population is able to be quickly retrained. A second family of solutions is based on rehearsing some previously learned items as new items are added to the network, but these methods require the separate storage of all previously learned information so that it is available for relearning.

Pseudorehearsal is very like rehearsal but does not require access to old information. Instead, pseudorehearsal approximates old information as needed by randomly sampling the behaviour of the network. Both methods work by forcing changes made to the function embodied by the network to be *local* to the new information being learned. Rehearsal / pseudorehearsal methods are related to other proposed solutions to the catastrophic forgetting problem in that they naturally result in a sharpening of hidden unit representations. (In contrast to other methods however, hidden units are also reordered, implying that considerable flexibility may be required to fully exploit sharpening).

The main significance of these methods is that they provide a practical way of extending the capabilities of current neural network learning algorithms to allow sequential learning (learning new information at any time). This should enable a range of topics, including the consolidation of newly learned information, ongoing / lifelong learning, developmental effects, and also transfer effects (see for example Robins (1997)), to be more easily modelled within the neural network framework.

In Robins (1996) we have also argued more specifically pseudorehearsal can be related to the "sleep consolidation" hypothesis. If the catastrophic forgetting problem has occurred during the evolution of the brain then a specific solution, a mechanism for consolidating knowledge, is obviously required. The sleep consolidation hypothesis proposes that newly learned information is consolidated into long term memory during sleep (see for example Winson (1990)). There are a number of similarities between pseudorehearsal and sleep consolidation. Both serve the function of consolidation without requiring explicit access to the old information (previous learning experiences) for relearning. Both involve the random stimulation of the "long term memory", pseudorehearsal by the construction of random pseudoitems, and sleep in the stimulation of the neocortex by random or chaotic input from the brainstem. Robins (1996) describes these similarities in more depth.

Extending rehearsal and pseudorehearsal methods to other network types has resulted in some interesting insights. In contrast to the feed forward "function approximation" networks described in this paper, Hopfield networks are recurrent "dynamical systems". As shown by Robins & McCallum (1998), rehearsal and pseudorehearsal (where pseudoitems are randomly chosen attractors in the network) are both effective in this context, but the distinction between the methods starts to break down. Randomly sampling the attractors of a network results in both novel "spurious" attractors and also actual attractors corresponding to learned items. Either can be effectively rehearsed to minimise catastrophic forgetting. We are currently exploring the relationship between this *relearning* effect, and the *unlearning* model of Crick & Mitchison (1983, 1986) as a model of the consolidation of information during sleep.

Acknowledgements

Sections of this paper are based on Robins (1996) and Frean & Robins (1997).

References

Ans,B. & Rousset,S. (1997) Avoiding Catastrophic Forgetting by Coupling Two Reverberating Neural Networks. *Academie des Sciences, Sciences de la vie, 320,* 989 - 997.

Burgess,N., Shapiro,J.L. & Moore, M.A. (1991) Neural Network Models of List Learning. *Network, 2,* 399-422.

Crick,F. & Mitchison,G. (1983) The Function of Dream Sleep. *Nature, 304,* 111-114.

Crick,F. & Mitchison,G. (1986) REM Sleep and Neural Nets. *Journal of Mind and Behaviourr, 7,* 229-249.

Frean, M. & Robins, A. (1997). *Catastrophic Forgetting in Neural Networks: Further Exploration of The Pseudorehearsal Solution and Implications for Consolidation and Transfer* (Tech. Rep. AIM-36-97-2). Otago, New Zealand: Otago University, Department of Computer Science.

French,R.M. (1992) Semi-distributed Representations and Catastrophic Forgetting in Connectionist Networks. *Connection Science, 4(3&4),* 365 - 377.

French,R.M. (1994) Dynamically Constraining Connectionist Networks to Produce Distributed, Orthogonal Representations to Reduce Catastrophic Interference. *Proceedings of the 16th Annual Cognitive Science Society Conference,* 335-340. Hillsdale, NJ: Earlbaum.

French,R.M. (1997) Interactive Connectionist Networks: An Approach to the "Sensitivity-Stability" Dilemma. *Connection Science, 9,* 353 - 380.

Grossberg,S. (1987) Competitive Learning: From Interactive Activation to Adaptive Resonance. *Cognitive Science, 11,* 23 - 63.

Hetherington,P.A. & Seidenberg,M.S. (1989) Is There "Catastrophic Interference" in Connectionist Networks? *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society,* 26 - 33. Hillsdale NJ: Lawrence Earlbaum.

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

Lewandowsky,S. (1991) Gradual Unlearning and Catastrophic Interference: A Comparison of Distributed Architectures. In Hockley,W.E. & Lewandowsky,S. (Eds.) *Relating Theory and Data: Essays on Human Memory in Honour of Bennet B. Murdok,* 445 - 476. Hillsdale NJ: Lawrence Earlbaum.

Lewandowsky,S. & Li,S. (1995) Catastrophic Interference in Neural Networks: Causes, Solutions, and Data. Dempster,F.N. & Brainerd,C. (Eds.) *Interference and Inhibition in Cognition.* San Diego: Academic Press.

McClelland,J.L., McNaughton,B.L. & O'Rielly,R.C. (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 (3),* 419 - 457.

McCloskey,M. & Cohen,N.J. (1989) Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In Bower,G.H. (Ed.) *The Psychology of Learning and Motivation: Volume 23,* 109 - 164. New York: Academic Press.

McRae,K. & Hetherington,P.A. (1993) Catastrophic Interference is Eliminated in Pretrained Networks. *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society,* 723 - 728. Hillsdale NJ: Lawrence Earlbaum.

Moody,J. (1994) Prediction Risk and Architecture Selection for Neural Networks. *In Cherkassky,V., Friedman,J.H. & Wechsler,H. (Eds.) From Statistics to Neural Networks: Theory and Pattern Recognition Applications.* NATO ASI Series F, Springer-Verlag.

Murphy,P.M. & Aha,D.W. (1994) UCI Repository of Machine Learning Databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

Murre,J M J (1992a) The Effects of Pattern Presentation on Interference in Backpropagation Networks *Proceedings of the 14th Annual Conference of the Cognitive Science Society* , 54-59. Hillsdale NJ: Earlbaum.

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

Nadal,J.P., Toulouse,G., Changeux,J.P. & Dehaene,S. (1986) Networks of Formal Neurons and Memory Palimpsets. *Europhysics Letters, 1,* 535 542.

Ratcliff,R. (1990) Connectionist Models of Recognition Memory: Constraints Imposed by Learning and Forgetting Functions. *Psychological Review, 97(2),* 285-308.

Robins,A. (1995) Catastrophic Forgetting, Rehearsal, and Pseudorehearsal. *Connection Science, 7,* 123 - 146.

Robins,A. (1996) *Consolidation in Neural Networks and in the Sleeping Brain. Connection Science, 8,* 259 - 275.

Robins,A. (1997) Towards a Framework for Consolidation and Transfer Effects in Neural Network Models. *Proceedings of The Fourth Australasian Cognitive Science Conference,* in press.

Robins, A. & McCallum, S. (1998). Pseudorehearsal and the Catastrophic Forgetting Solution in Hopfield Type Networks *Connection Science,* in press.

Rumelhart,D.E. Hinton,G.E. & Williams,R.J. (1986) Learning Internal Representations by Error Propagation. In Rumelhart,D.E., McClelland,J.L. & the PDP Research Group (1986) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations. Cambridge MA: MIT Press.

Sharkey,N. & Sharkey,A. (1995) An Analysis of Catastrophic Interference *Connection Science, 7,* 301 - 329

Winson,J. (1990) The Meaning of Dreams. *Scientific American, November 1990,* 42 - 48.