Connection Science, Vol. 13, No. 3, 2001, 257–275

Retroactive interference in neural networks and in humans: the effect of pattern-based learning

DANIEL MIRMAN* and MICHAEL SPIVEY†

**Department of Psychology and The Centre for the Neural Basis of Cognition, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213–3890, USA*
email: dmirman@andrew.cmu.edu
tel: +1 412 268-8112
fax: +1 412 268-2798

†*Department of Psychology, Cornell University, USA*
email: spivey@cornell.edu
tel: +1 607 255-9365
fax: +1 607 255-8433

Abstract. Catastrophic interference is addressed as a problem that arises from pattern-based learning algorithms. As such, it is not limited to artificial neural networks but can be demonstrated in human subjects in so far as they use a pattern-based learning strategy. The experiment tests retroactive interference in humans learning lists of consonant–vowel–consonant nonsense syllable pairs. Results show significantly more interference for subjects learning patterned lists than subjects learning arbitrarily paired lists. To examine how different learning strategies depend on the structure of the learning task, a mixture-of-experts neural network model is presented. The results show how these strategies may interact to give rise to the results seen in the human data.

Keywords: catastrophic interference, neural networks, mixture of experts, memory.

1. Introduction

1.1 *Catastrophic interference in neural networks*

McCloskey and Cohen (1989) and Ratcliff (1990) were the first to investigate thoroughly catastrophic interference (CI) in neural networks. They found that when feedforward neural networks using the backpropagation learning algorithm (Rumelhart *et al.* 1986) were trained sequentially they exhibited a 'catastrophic' level of interference. Training in these networks consists of presenting inputs, calculating the error function (difference between network output and target output) and performing a gradient descent to minimize the error function. Sequential training refers to completely training (until error is reduced below a criterion) the network on an input–output set and then training on another set. This type of training contrasts with interleaved training in which items from both lists are mixed together during training. Importantly, no significant interference effects are seen during interleaved training, indicating that CI is limited to sequential learning.

Connection Science ISSN 0954-0091 print/ISSN 1360-0494 online © 2001 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/0954009011009183 0

Several researchers have developed learning algorithms that alleviate the problem of CI in sequential learning (see French 1999 for a recent review). In general, these models strive to reduce the extent to which the hidden layer representations are distributed (French 1992, Kruschke 1992, Sharkey and Sharkey 1995). Kruschke's (1992) ALCOVE model was able to avoid CI by employing a localization algorithm for hidden node activation:

$$a_j = \exp(-c\sum \alpha_i \left| h_{ji} - a_i \right|)$$

where a_j is the jth hidden node activation, a_i is the activation of the ith input node, α_i is the attentional weight, c is a hidden node specificity parameter and h_{ji} is the ith coordinate for the position of the jth hidden node. For each hidden node there is a vector, which can be thought of as a set of coordinates for the position of the hidden node. This 'position' specification allows a hidden node to be localized to a specific input pattern. ALCOVE is a special case of a general class of networks known as radial-basis function networks (Poggio and Girosi 1990). This type of network has hidden nodes whose activation is a function (typically a Gaussian) of the distance between hidden node location and the input vector. In Kruschke's model there was one hidden node for each input pattern and each hidden node had a 'position' vector corresponding to one of the input patterns. Thus, it could be ensured that only one hidden node would be active for each input pattern and (with appropriate α and c values) that that node's activation would be 1, while all the others would be 0. This model reduces the distributed nature of the hidden layer's representation of the inputs; it localizes it, and in the extreme makes the representations perfectly non-overlapping.

It is important to note that while extreme localization of the input patterns will remove CI it will also remove some of the most interesting and useful properties of these types of neural networks, namely, generalization and representational efficiency. A neural network utilizing non-overlapping representations will have no representation of functional similarities between inputs. For example, if such a network is trained to categorize robins and canaries as 'birds', and oaks and pines as 'trees', there will be no more similarity between the representations for robin and canary than for robin and oak because each of the representations will have no overlap with the others (McClelland *et al*. 1995). Furthermore, suppose that the network has been trained to know that both robins and canaries, aside from being birds, have wings and can fly. Then the network is presented with a novel item such as a sparrow and trained to know that a sparrow is a bird. A network using overlapping, distributed representations will be able to guess that sparrows have wings and can fly, however, a network using non-overlapping representations will not.

French (1992) noted this trade-off between reducing CI and reducing a network's ability to generalize. His solution was a semi-distributed network that strives for a balance between the extremes. He used a node sharpening technique that imposed fewer restrictions than Kruschke's localizations but accomplished a similar reduction in overlap of hidden node representations. For each input pattern, a hidden node representation was calculated using standard feedforward procedures, then this activation was used to produce target hidden node activations which were less distributed. This target was produced by 'sharpening' the activation for the k most active hidden nodes according to the formulae:

$$A_{new} = A_{old} + \alpha(1 - A_{old}) \quad \text{for the } k \text{ nodes to be sharpened}$$
$$A_{new} = A_{old} - \alpha A_{old} \quad \text{for the other nodes}$$

where α is the sharpening factor. The difference between the target activation and the actual activation was used to adjust the input to hidden layer weights. Then the input pattern was fed forward from input to output and all the weights were adjusted according to standard backpropagation of errors (Rumelhart *et al*. 1986).

To analyse the performance of this type of learning algorithm, French trained a model with eight input, eight hidden and eight output nodes on a set of 11 associations. After this training session, the network was presented with a new association, and after it had learned the new association it was tested on an association from the first set. The model generally performed poorly if output was compared to the target output. However, a more sensitive measure of forgetting was examined: the number of presentations required to relearn the association. In this evaluation a strong relationship between the number of nodes sharpened and the amount of forgetting was found. A standard backpropagation network without node sharpening required more than four times as many presentations to relearn the association as a one-node sharpened network, and the two-node sharpened network performed even slightly better. If more nodes were sharpened, the number of presentations required to relearn began to increase until with four- and five-node sharpening the network was no better than the standard. To evaluate the original hypothesis, hidden node activation overlap was examined and was found to be at a minimum for one-node sharpening, and to increase steadily, reaching the standard level at four-node sharpening.

French then used real-world data in the form of voting records to test the algorithm. A network was designed to categorize inputs into one of two categories (Democrat/Republican) based on voting records from 16 issues (a yes vote was coded as a 1 and a no as a 0 in the input vector). The network showed improvement in relearning time (although the improvement was only moderate compared with the earlier experiment) and showed high generalization ability that was independent of the number of nodes sharpened.

Sharkey and Sharkey (1995) analysed the problem of CI and concluded that the solution lies in localization of hidden node representations. Their proposed solution is a generalization of the previously discussed models. The HARM model divides the learning task into two subtasks. The first subtask is to eliminate overlap in the input patterns such that each one is mapped to a unique hidden node. The second subtask is to produce the appropriate output from the hidden nodes. Since the hidden node activation patterns are non-overlapping, this task can be accomplished with a single weight matrix and Hebbian learning. This general model design can be applied to many different learning algorithms since the effects of the functions are specified and not the mathematical details. That is, any localizing learning rule can be used in the first step. This includes Kruschke's (1992) node localization and French's (1992) node sharpening algorithms that produce orthogonal hidden node representations if appropriate parameters are used (Kruschke: α=l, c=l0; French: k=l).

One weakness of this model is that for an input vector of length n, 2^n hidden nodes are required. Since localization is the extreme of reduction of hidden node overlap, any model that strives to reduce distribution of hidden node representations would face this constraint in eliminating CI. Of course, if the number of input patterns is known, only that many hidden nodes are required, but for some applications this is not a reasonable constraint. Furthermore, this sort of localizing-associating model of

human memory may seem unsatisfactory to cognitive psychologists who consider human memory to be more complex than a simple look-up table. That is, human memory exhibits properties that are not well modelled by a look-up table, such as generalization (Shepard 1987), rule abstraction (Simon and Kotovsky 1963, Restle and Brown 1970), hierarchical organization of sequences (Jones 1974, Deutsch and Feroe 1981) and chunking (Miller 1956). None the less, it should be noted that researchers such as Hintzman (1984) have argued in favour of look-up table models of memory and implemented them to model successfully many aspects of human memory.

Some researchers have attempted to solve the problem of CI with models based on learning strategies used by humans. McRae and Hetherington (1993) argue that humans do not undertake new learning tasks with randomly set weights; instead we bring a wealth of previous knowledge to a task and this helps us avoid large interference. To simulate this previous knowledge, the researchers pretrained a network on patterns similar to the patterns on which it was to be tested. The pretraining technique was tried on a small network using simple pattern learning and on a large network using a set of 2897 patterns from Seidenberg and McClelland's (1989) model of word naming. This pretraining was thought to provide the network with something comparable to knowledge of the English language. The model was then tested on CVCs (consonant–vowel–consonant sets) constructed following the same rule as the words in the pretraining corpus. The performance of pretrained and naïve networks on three different sequential learning tasks was compared.

It was found that pretrained networks suffered from no interference during these tasks. The error increase above training criterion was six times greater for naïve networks than for pretrained ones. The activation of the hidden layer was also examined and it was found that pretrained networks used fewer hidden nodes to represent each input pattern, meaning that the representations were less distributed. Thus, while these researchers did not set out to localize the hidden layer representations, they ended up with networks that were moving in that direction. It is important to note that although the representations were less distributed, they were still partially overlapping, yet the results were better than many previous experiments. This suggests that representational overlap may be an imperfect measure of hidden node overlap and a more accurate measure should be employed in analysing the causes of CI.

Rehearsal is another approach derived from studying human learning. Ratcliff (1990) used a rehearsal buffer consisting of three previously learned items and one new item to train his network. Once this set had been learned to criterion, one of the items was replaced with a new item. This process continued such that the training set always consisted of the three most recently learned items and one new item. This technique improved performance by a small amount and was not thought to be very significant.

More recently, Robins (1995) conducted more extensive simulations using different rehearsal algorithms and found that some were much more effective than Ratcliff's (1990) recency rehearsal. In these simulations, three previously learned items were trained concurrently with the new item, but the way those three items were selected from the trained corpus was varied. Random rehearsal refers to a random selection of items from the trained corpus and those items are trained with the new item for the number of epochs required to reach criterion. Sweep rehearsal uses a 'dynamic' training buffer: for each training epoch three items are randomly selected from the trained corpus for training with the new item; for the next epoch three new items are

selected from the trained corpus. In comparing these two algorithms it can be said that random rehearsal provides 'narrow and deep' training, while sweep rehearsal provides 'broad and shallow' training. The results were impressive—random rehearsal showed only small interference and sweep rehearsal showed no interference and actually improved performance on some items while new ones were being learned.

Even with this effective form of rehearsal, the algorithm relies on the network having access to the entire trained corpus at all times. For many applications, including the modelling of human memory, this is not reasonable. Robins (1995) also tested a technique he called 'pseudorehearsal'. Pseudorehearsal assumes that the network does not have access to the items it has already learned and so it must produce pseudo-items. Pseudo-items are produced by generating a random input vector of ones and zeros and feeding it through the network. The output becomes the associated target vector. This pseudo-population acts as a map of weights in the network before training on a new item begins. The size of the pseudo-population was found to have a significant effect on the effectiveness of sweep rehearsal, with bigger populations producing less interference. Both sweep and random rehearsal were more effective than pseudorehearsal, but with sufficiently large pseudo-populations, sweep pseudorehearsal showed only a small amount of interference. These simulations show that examination of human learning behaviour can be as effective at improving connectionist models as studying their mathematical mechanics.

1.2 *Retroactive interference in humans*

In humans, the forgetting of previously learned information after new learning has been termed retroactive interference (RI) and is very small compared to the CI of neural networks. Barnes and Underwood (1959) demonstrated RI using the AB–AC paradigm: subjects were trained to respond with adjective B to nonsense syllable A (from Glaze 1928) until they learned eight AB pairs. Then they were trained to respond with a different adjective C to the same nonsense syllable A. The number of AC list learning trials was varied (1, 5, 10, 20). Then the subjects were asked to give both the B and C adjectives in response to A. A–B list performance was seen to decrease as the number of A–C list learning trials increased, but even at the maximum (20) the performance was near 50%—well above CI performance, which is typically near 0% (McCloskey and Cohen 1989, McClelland *et al*. 1995).

McClelland *et al*. (1995) argue that it is the complementary learning systems of hippocampal structures and the neocortex that prevent CI in humans. This argument is motivated by neurobiological evidence and supported by connectionist models. They suggest that at first sequential learning is handled by more or less localized representations in the hippocampus, which then trains the neocortex over a long period of time with new as well as old items so that the neocortex can form distributed representations of the entire corpus of knowledge. French (1997) and Ans and Rousset (2000) have also developed connectionist models of complementary learning systems incorporating the technique of pseudorehearsal (Robins 1995). However, these accounts may not address short-term learning tasks of the type studied by Barnes and Underwood (1959).

Although CI in tasks often modelled with neural networks has not been demonstrated in biological systems, a few researchers have found effects that are similar in rat perception of time duration (French and Ferrara 1999) and human motor memory (Shadmehr and Brashers-Krug 1997). French and Ferrara (1999) trained rats

to expect food pellets at 40-s intervals and at 8-s intervals. If the rats were trained first on one duration and then on the other they continued to expect the food pellet at the latter duration. Conversely, if the rats were trained on both durations concurrently (i.e. interleaved training) their response was clearly bimodal, showing expectation at both time intervals. Thus, sequential training led to memory of the last set only, while interleaved training led to learning of both sets.

Shadmehr and Brashers-Krug (1997) found in humans that if training sessions with two conflicting mechanical environments were not separated by about 5 h, the learning of the second would 'overwrite' the first: learning of the second would take longer (indicating that learning started with a representation of the first rather than a 'tabula rasa') and would interfere with recall of the first. This is analogous to CI in that CI results from encoding new representations over previously learned representations, producing large interference effects.

1.3 *Hierarchical modular connectionist architectures*

Some aspects of human cognition seem to rely on modular processing and some researchers have attempted to take advantage of modular neural processing to develop more robust and accurate models. In particular, models of multispeaker phoneme recognition (Hampshire and Waibel 1992) and 'what-where' vision tasks (Jacobs *et al*. 1991, improved algorithm in Jordan and Jacobs 1994) have been improved by the use of modular connectionist architectures. Modular connectionist architectures are connectionist models that have sub-networks which compete to represent input–output pairs. A gating network outputs a coefficient vector; these coefficients are used by the output layer to combine the sub-network outputs to make a final output. In the extreme competitive case the coefficient vector has one 1 and all the other values are 0, in a more co-operative model the values can vary so that the final output can be a weighted sum of the sub-network outputs.

The finding of Jacobs *et al*. (1991) that is most relevant here is that a modular connectionist network with architecturally different sub-networks will partition the learning task in a way that takes advantage of the different properties of the architectures of the sub-networks. The input was a single vector of 25 pixel values and a task specifier; however, the network divided what and where processing to different sub-networks. Thus, the network learned to dedicate its separate modules to 'what' and 'where' vision tasks, analogous to ventral and dorsal visual pathways in the primate brain (Mishkin *et al*. 1983).

2. Experiment

Let us assume that humans are capable of at least two kinds of learning strategies: pattern-based (Restle and Brown 1970, Jones 1974) and rote memorization (Hintzman 1984). This assumption seems intuitively correct in that there are some learning tasks that require arbitrary associations (such as face-first name association) and some that follow set rules (driving a car). These two strategies differ along two dimensions: generalization and resource limitations. Pattern-based learning is relatively resource efficient and allows for generalization to similar situations (e.g. once a person learns to drive a car, that person can drive any similar car); on the other hand, rote memorization requires more resources and does not allow generalization (e.g. people that look similar do not necessarily have the same first name). These

differences are analogous to differences between neural networks using distributed and localized representations, with distributed representations being the analogue of pattern-based learning, and localized representations that of rote memorization.

Three statements can be made based on the described studies and the above assumption:

(1) The subjects in Barnes and Underwood's study (1959) used rote memorization and suffered only minor interference analogous to the small amount of interference suffered by networks using learning algorithms that orthogonalize hidden node representations (French 1992, Kruschke 1992, Sharkey and Sharkey 1995).
(2) An experiment that presents subjects with a learning task suitable for pattern-based learning should produce increased interference.
(3) A modular connectionist architecture should be able to model the interaction of these two learning strategies.

The final two propositions are the goal of the present study: this section addresses statement (2) the Simulation section addresses statement (3).

2.1 *Methods*

2.1.1 *Participants.* The subjects were 32 students at Cornell University who participated in exchange for extra credit in Psychology and Human Development courses.

2.1.2 *Stimuli and procedure.* Instructions, stimulus presentation and testing were conducted on computer using PsyScope software (Cohen *et al.* 1993). Each subject was shown list A, containing 10 nonsense syllable (CVC) pairs chosen to minimize associative value according to Glaze (1928) and to remain pronounceable. To maximize similarity to neural network training, the 10 pairs were presented one at a time for 5 s each, cycling through the entire list four times. The entire training time for list A (including 700 ms pauses between CVC pairs) was about 4 min per list. Then the subjects were shown the first member of each pair and asked to choose the correct match from two possible choices. Although this forced-choice paradigm does not exactly reflect the network's task, it allows the greater advantage of knowing the exact chance performance so that the data are easier to interpret. This first test contained 18 probe items with two choices for each one. The additional eight test items were used to test for generalization, and to match the second test. Items from the learned list and generalization items were mixed randomly on the test. Subjects were instructed to skip syllables that they did not recognize (both in the instructions displayed before beginning the experiment and verbally in the case that they asked the experimenter). The learning procedure was then duplicated with list B, which contained 10 new CVC pairs. Finally, subjects were tested on both lists, using a test of exactly the same format as the first test. This test also contained 18 items but the extra items were from the first list rather than generalization probes.

Subjects in the control condition were shown lists consisting of arbitrarily paired CVCs. Presumably learning of this list would require rote memorization and would not be as susceptible to interference as the pattern-based experimental lists. The experimental lists consisted of CVCs paired according to a simple rule: on one list the

CVCs followed an inversion pattern (VEC–CEV), on the other list they followed a vowel exchange pattern (KIZ–KEZ, GAK–GOK) (see the Appendix).[1] One half of the experimental subjects received the inversion list first and the vowel exchange list second, the other half received them in the opposite order. During control condition testing, an item would be presented with the correct response and a random distracter CVC, and the subject chose one alternative. During experimental condition testing, an item would be presented with the appropriate alternatives according to the two different patterns, and the subject chose one alternative. Thus, the experimental test task was reduced to identifying the source list of the presented CVC and applying the corresponding pattern.

2.2 *Results*

On the first test, subjects in the pattern-based learning condition made fewer errors than control subjects (experimental: 0% pre-interference error; control: 5.3% pre-interference error). However, after the second list (interference list) was learned, they made more errors than control (experimental: 17.0% post-interference error; control: 8.2% post-interference error; figure 1). A main effect of list was found ($F(1,30)=18.16$, $p<0.001$) confirming that, overall, subjects showed decreased performance (retroactive interference) after the second list was learned. Most importantly, the condition X list interaction corresponding to a difference in interference effects was found to be significant ($F(1,30) =9.07$, $p<0.01$).

Additionally, it was found that, on the first test phase, subjects learning pattern-based lists attempted to generalize more frequently than subjects learning lists without patterns (experimental: 6.3 mean generalizations; control: 3.6 mean generalizations; $t(30) = 2.63$, $p < 0.02$; figure 2) and, of course, did so much more accurately (experimental: 6.2 mean correct generalizations; control: 1.8 mean correct generalizations; $t(30) = 5.56$, $p < 0.0001$; figure 2). For the experimental condition, a generalization was

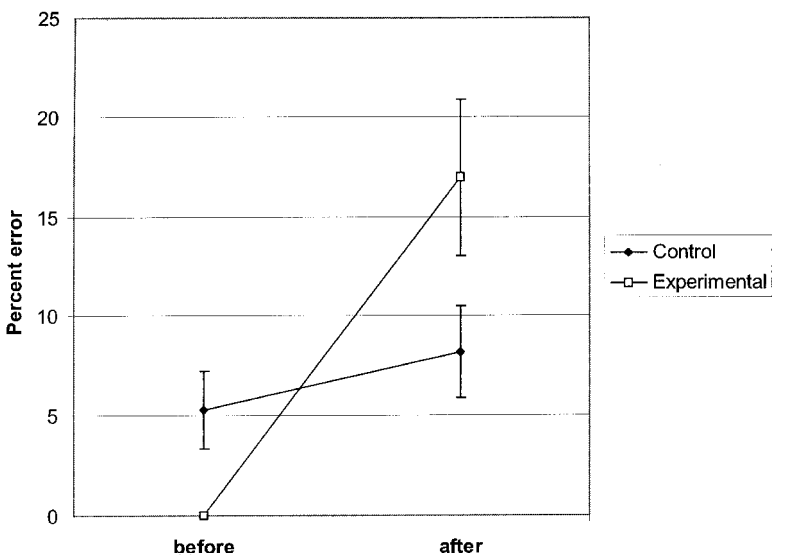

Figure 1. Effect of interference on per cent error by human subjects by condition. Subjects in the experimental condition show increased interference over controls.

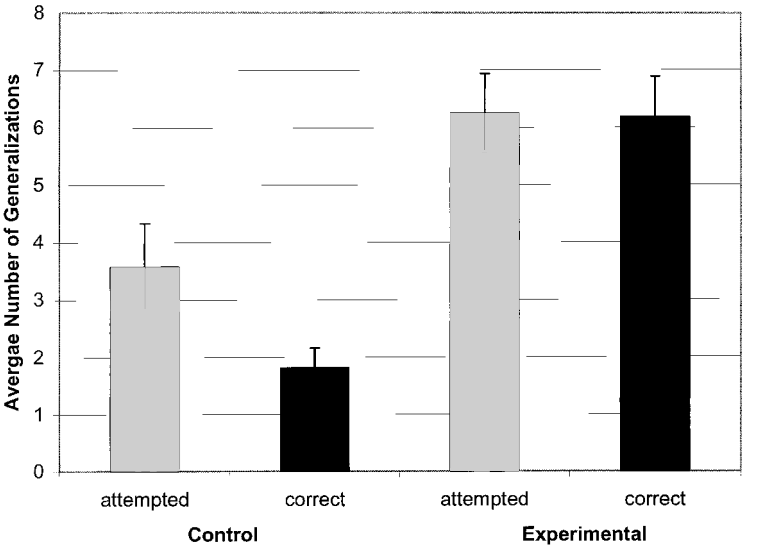

Figure 2. Average number of attempted and correct generalizations for human subjects. Subjects in the experimental condition generalize more frequently and more accurately.

considered correct if the response followed the pattern; for the control condition a response was arbitrarily assigned as correct. Obviously, attempts to generalize in the control lists were arbitrary, and accuracy was at chance. Since subjects were instructed not to generalize, that they did generalize suggests that they were encoding a general mapping rather than specific pairs from the list.

2.3 *Discussion*

This experiment showed that human subjects learning lists with simple patterns were susceptible to greater retroactive interference than subjects learning lists of arbitrarily paired CVCs. Two findings suggest that the experimental subjects used a learning strategy resembling pattern-based learning, while control subjects used a rote memorization strategy:

(1) Patterned list subjects generalized much more than control subjects.
(2) Every patterned list subject responded correctly to 100% of first list items before interference, while control subjects averaged 94.7% correct ($F(1,30) = 7.47$, $p < 0.05$) and showed much greater variability.

If it is true that experimental subjects used a pattern-based learning strategy while control subjects used a rote memorization strategy then the increased retroactive interference (figure 1) can be seen as resembling CI in neural networks. However, considering that 50% correct would be chance performance on this task, the 83% post-interference correct responses from experimental subjects is far from the results of CI where performance typically drops to chance. The most likely explanation is that experimental subjects did not use fully distributed and overlapping representations but rather semi-distributed representations. As shown by a number of investigators (French 1992, Kruschke 1992, Sharkey and Sharkey 1995), representational overlap

forms a continuum from fully localized (only one node active in representing each input vector) to fully distributed (every node active in representing each input vector), and CI varies along this continuum in direct proportion to the amount of overlap. The most likely conclusion, then, is that experimental subjects fell on this continuum significantly closer to distributed representations than control subjects and thus showed increased CI. We suggest that this continuum is valid for humans as well as artificial neural networks and that—depending on the structure of the learning task—human subjects will exhibit learning and forgetting behaviour that varies along this continuum. This means that connectionist networks are valuable tools for modelling human memory despite differences in interference behaviour and, in particular, modular architectures that allow for selection/competition among different learning strategies appropriate to the structure of the learning task may be appropriate models of human learning and memory.

3. Simulation

3.1 *The model*

Along the lines of the mixture-of-experts networks (e.g. Jacobs *et al.* 1991), the Dual-Strategy Competitive Learner (DSCL) is composed of two sub-networks and a gating network. Each of the sub-networks represents an 'expert' in that it has a unique architecture and learning algorithm making it differentially effective based on the learning task. The gating network is trained to decide which expert is the correct one for a given input; that is, which sub-network's output will be used as the overall output (figure 3).

The input layer has 15 units. One of the sub-networks is a standard backpropagation network (Rumelhart *et al.* 1986) with 10 hidden units. The other sub-network is an implementation of ALCOVE (Kruschke 1992) with 20 hidden units, each dedicated to one of the 20 input patterns (10 from each list). The gating network is also a backpropagation network with 10 hidden nodes which outputs a two-element bit vector designating which sub-network will be used as network output (i.e. [1 0] for expert 1, [0 1] for expert 2). Network constants are given in table 1.

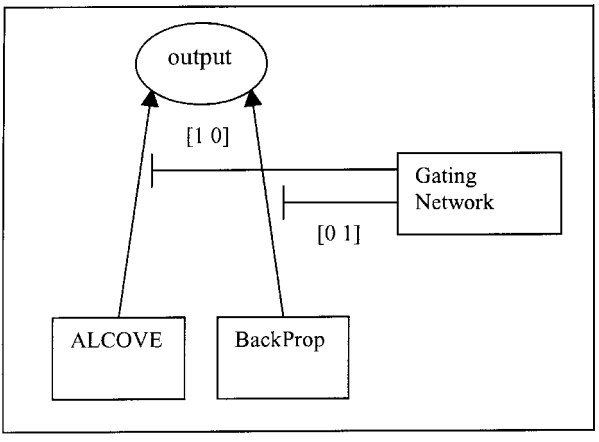

Figure 3. Modular architecture of DSCL. Output of either backpropagation or ALCOVE sub-network is selected by gating network as DSCL output.

Table 1. Network constants used in DSCL.

BackProp expert learning rate	0.2
BackProp expert momentum	0.9
ALCOVE expert attention learning rate	0.001
ALCOVE expert association learning rate	0.3
ALCOVE expert specificity constant	2
Gating network learning rate	0.1
Gating network momentum	0.9

Table 2. Some examples of inputs and targets used to test DSCL.[a]

List name	Input	Target	Per cent pairs won by ALCOVE
Control	000100010010000 100000100000001 010000010000010	100001000010000 001000010000001 001000010001000	75.2
Auto-associative	111111000000000 011111100000000 001111110000000	111111000000000 011111100000000 001111110000000	8.6
Inverse	101010000000000 010101000000000 001010100000000	010101111111111 101010111111111 110101011111111	0.0

[a] The right column shows the per cent of pairs won by ALCOVE. Control (arbitrarily paired) lists are won mostly by ALCOVE, the patterned lists are won almost entirely by the backpropagation expert.

Each list of inputs consists of 10 15-element bit vectors. The target outputs are also 15-element bit vectors. Control lists are arbitrarily paired bit vectors while experimental lists followed a pattern, with a different pattern represented in the first list and the interference list. An auto-associative pattern (each target is the same as its input) was used for list A and a 0–1 inversion pattern (0s become 1s and vice versa) for list B (see table 2 for examples of input–target pairs). In designing the CVC experimental lists patterns were chosen that subjects found easy to recognize; similarly, in designing the experimental inputs for the network patterns were chosen that would be easy for the network to encode.

Training consisted of presenting an item in the list to all three component networks, calculating error for each sub-network and adjusting weights according to the appropriate learning algorithm in the expert sub-network which showed least error—this expert was called the winner. Finally, the gating network was trained to choose the winner's output as the overall network output for this input vector. This procedure was repeated for each vector in the training list. In this manner the list was presented 30 times (i.e. 30 epochs). For comparison, a standard backpropagation network was trained on the same input/target patterns using 30 epochs and 100 epochs.

First the network was trained with list A, then it was tested with list A (to measure pre-interference performance) and with a different, generalization list. For the control

case, the generalization list also had arbitrarily paired inputs and targets. For the patterned case, the generalization list input and target vectors followed the same pattern as the training list. In this way it was possible to test whether the pattern learned during training could be generalized to novel vectors in the same way that human subjects generalized patterns learned during training to novel CVCs. Following these tests, the network was trained with list B and tested with list B and list A (to measure post-interference performance).

The DSCL was initialized to a random state before starting control or experimental training. This was done so that strategy choice would be determined by the network itself (in the form of winner selection by the gating network) based on the type of input it saw; just as the subjects chose their strategies based on the type of stimuli they saw.

3.2 *Simulation result*

Network performance was measured primarily by per cent of inputs for which the network would generate the incorrect output vector. An output was judged as correct only if each node's activation was on the target side of 0.5 (i.e. less than 0.5 for a target of 0, greater than 0.5 for a target of 1.0). However, with 15-bit output vectors this is a very difficult task (chance performance being about 0.003%, compared to 50% for the human subjects' task), so in some cases it was useful to look at mean squared error across the input patterns.

Reported DSCL results are averages of at least 40 runs (figure 4). Performance was very similar to humans—showing much larger interference in patterned lists than control (arbitrarily paired) lists. An examination of the gating network outputs (winners, figure 5) shows that ALCOVE won most of the control (unpatterned) pairs and the backpropagation network won most of the patterned pairs.[2]

A backpropagation network trained by itself, for comparison, did not show any difference in amount of interference between control and patterned lists when per cent correct was used as a performance measure (figure 6), but it did show a difference when the more sensitive mean squared error performance measure was used

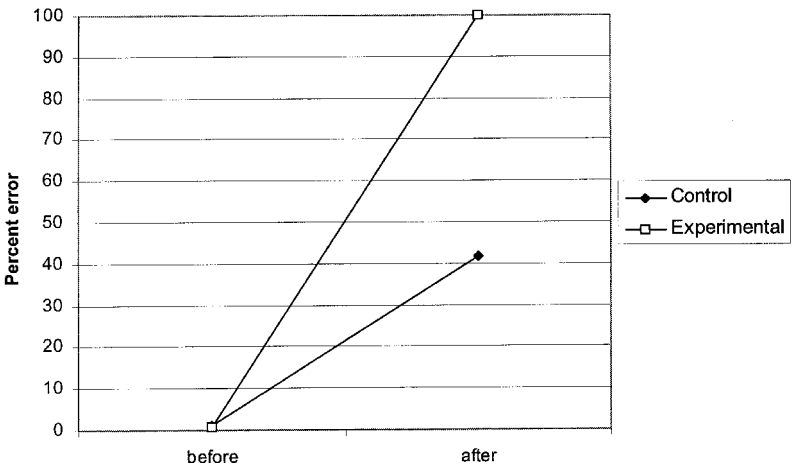

Figure 4. DSCL performance measured by per cent error. The experimental (patterned pairs) condition shows more interference compared to control (arbitrary pairs).

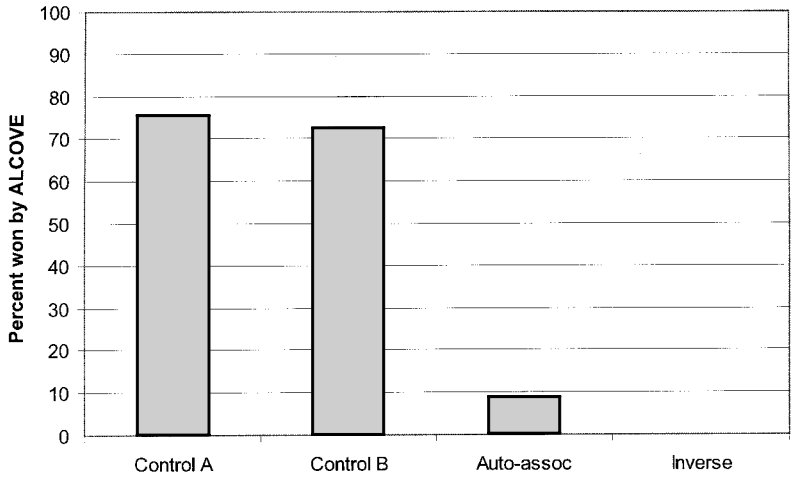

Figure 5. Per cent of each list's items won by ALCOVE. The control lists are won mostly by ALCOVE, the pattern lists are not (they are won by the backpropagation network).

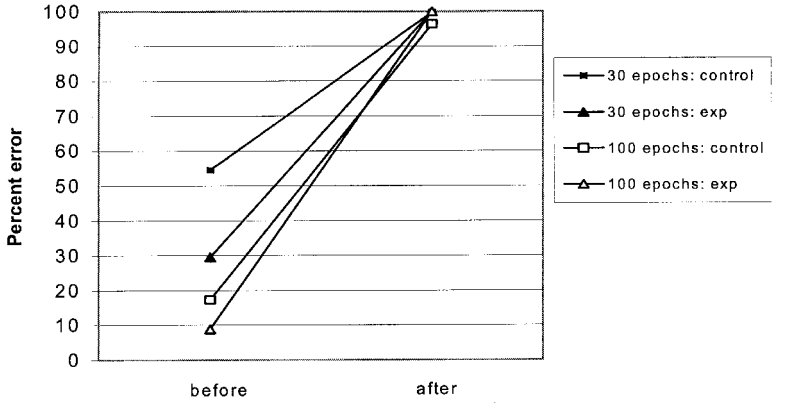

Figure 6. Performance of standard backpropagation network on control and experimental lists measured by per cent error. Post-interference performance is the same across lists.

(figure 7). However, even after 100 training epochs the difference did not show up in per cent correct.

Generalization results were supportive in modelling human behaviour by showing increased accuracy on experimental lists over control lists. The effect was particularly noticeable if mean squared error was used as the performance measure (figure 8), per cent correct showed only a very small difference between control and patterned list generalization ability (figure 9). The standard backpropagation network performed almost as well on the generalization tasks, which is not surprising since it is the backpropagation expert in DSCL that is responsible for generalization.

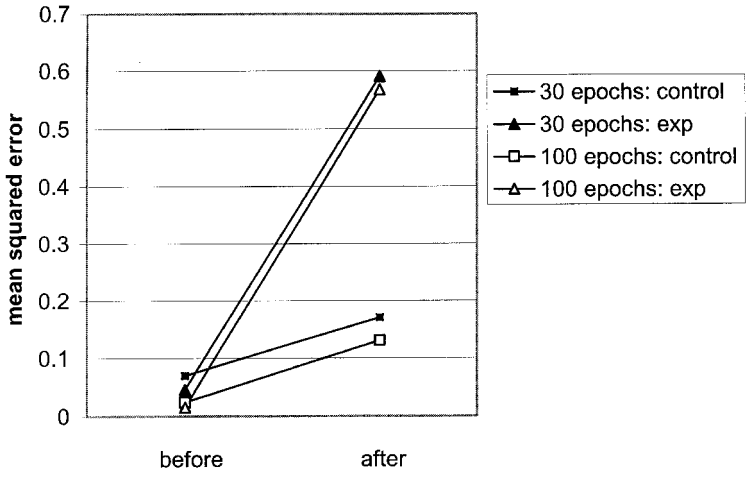

Figure 7. Performance of standard backpropagation network on control and experimental lists measured by mean squared error. Post-interference performance is clustered by list type with control lists having lower error.

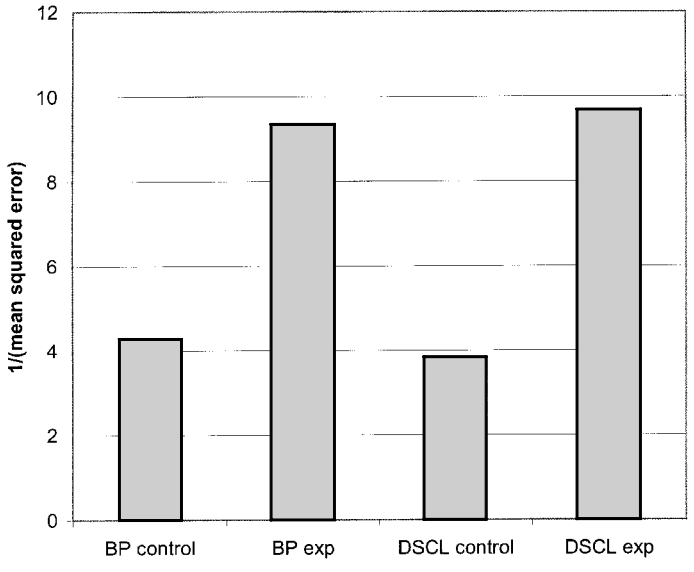

Figure 8. Generalization performance measured by l/(mean squared error). Performance is higher for the experimental (patterned) case.

3.3 *Discussion*

Network performance simulated human performance in showing increased retroactive interference for patterned lists over arbitrarily paired lists. In a recent study, Erickson and Kruschke (1998) modelled categorization using a modular architecture consisting of a rule module, an exemplar module and a competitive gating mechanism. They found that this model (called ATRIUM) accurately captured the interaction

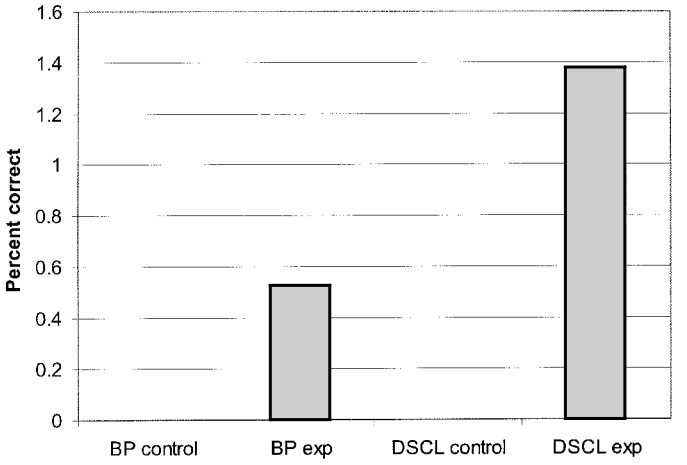

Figure 9. Generalization performance measured by per cent correct. Performance is higher for the experimental (patterned) case.

of rule- and exemplar-based categorization behaviour. This model bears a strong architectural resemblance to the model presented here for modelling human paired association learning and interference. However, there are two crucial differences: Erickson and Kruschke's (1998) human experiments present the rules explicitly, thus they use a rule module that explicitly represents the rule. In the present study human subjects extracted the patterns from the lists of CVCs without being instructed to do so. Thus, the model is designed to and succeeds in extracting patterns from the input–target vector pairing without explicit rule presentation. In addition, the ATRIUM gating mechanism takes input from the exemplar module (ALCOVE) and returns an output that represents the extent to which the exemplar module is the one used, that is, the final ATRIUM output is a linear combination of the outputs of the rule and exemplar modules. Conversely, the DSCL gating mechanism takes the same input as the sub-networks and outputs binary values such that the final network output is the output of one of the sub-networks, not a combination of them.

4. General discussion and conclusions

It has been suggested that because certain connectionist networks suffer catastrophic interference they are not viable models of human learning and memory (McCloskey and Cohen 1989). Interpreting CI as a result of pattern-based learning led to a retroactive interference experiment in association of CVC pairs. The experiment showed that human subjects exhibit significantly greater interference when the CVCs are paired according to an obvious pattern. Initially, this suggests that CI may not be grounds for discounting these types of networks as viable models of human memory. The results are also consistent with the hypothesis that humans use different learning strategies depending on the material to be learned. In particular, humans use a pattern-based learning strategy when it is appropriate and this strategy is more susceptible to retroactive interference.

An extension of pattern-based learning is the ability to generalize, which is not afforded by rote memorization (the alternative learning strategy in this experiment).

Human subjects learning patterned lists were found to generalize more frequently and more accurately than control subjects.

The interplay of the two learning strategies was modelled by DSCL, a competitive hierarchical modular neural network using backpropagation (Rumelhart *et al.* 1986) to model pattern-based learning and ALCOVE (Kruschke 1992) to model rote memorization. Per cent correct performance of the network approximated human performance. In addition, the network modelled human generalization results by showing less generalization error on patterned lists.

The results of the human experiment raise a few interesting questions and suggest possible follow-up experiments. (1) Is the choice of learning strategy under subjective control? That is, would it be possible to get the same results by presenting subjects with patterned lists and instructing one group to use a pattern-based strategy while instructing the other group to ignore the pattern? (2) How robust is the pattern-learning strategy? Would it work for more subtle patterns or imperfect patterns (patterns with exceptions)? Could it be that there is a continuum of 'patternhood', and the extent of pattern-based learning (and, consequently, the extent of interference) varies along it? (3) Is this effect specific to language or is it possible to replicate these results using non-language stimuli—pairing shapes and colours, for example?

This study shows that human memory is not easily described as a single simple process—in some cases it looks like a look-up table or rote memorization, while under different circumstances it looks like pattern extraction or function approximation. It is found that catastrophic interference is not limited to connectionist architectures but is a general product of pattern-based learning that occurs in humans as well. It can be seen in neural networks that perform function approximation and in humans when they use a pattern-based learning strategy rather than a rote memorization strategy, as is studied in most memory experiments (e.g. Barnes and Underwood 1959). These two learning strategies can be modelled by two different artificial neural network architectures, which account for a number of effects, including increased interference and increased generalization ability in pattern-based learning. These different neural network architectures can be put together in a competitive modular super-network that simulates the two different modes of human learning examined in this experiment.

Acknowledgements

We thank Edward Munandar for his help in developing and testing the model, Melinda Tyler for helping set up the experiment and reading an early draft, and Brent Vander Wyk, Shimon Edelman and David Field and two anonymous reviewers for their many insights and helpful comments and suggestions.

Notes

1. Some partial overlap in the items from the two experimental lists (e.g. GOW in the inversion list and GOK in the vowel switch list) was unavoidable because of the limited set of nonsense CVCs with low associative value. However, this was preferable to using the AB–AC training format of Barnes and Underwood (1959) because it more closely approximates the typical neural network training procedure.
2. Interestingly, some patterned lists were not recognized as such if they were too sparse (e.g. only three units active in a 15-node vector). Presumably, backpropagation cannot recognize patterns if they are too sparse.

References

Ans, B., and Rousset, S., 2000, Neural networks with a self-refreshing memory: knowledge transfer in sequential learning tasks without catastrophic forgetting. *Connection Science*, **12** (1): 1–19.

Barnes, J. M., and Underwood, B. J., 1959, Fate of first-list associations in transfer theory. *Journal of Experimental Psychology* **58**: 97–105.

Cohen, J. D., MacWhinney, B., Flatt, M., and Provost, J., 1993, PsyScope: a new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers*, **25** (2): 257–271.

Deutsch, D., and Feroe, J., 1981, The internal representation of pitch sequences in tonal music. *Psychological Review*, **88** (6:)503-522.

Erickson, and Kruschke, J. K. 1998, Rules and examplars in category learning. *Journal of Experimental Psychology: General*, **127** (2): 107–140.

French, R. M., 1992, Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science*, **4**: 365–377.

French, R. M., 1997, Pseudo-recurrent connectionist networks: an approach to the 'sensitivity-stability' dilemma. *Connection Science*, **9** (4): 353–379.

French, R. M., 1999, Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, **3** (4): 128–135.

French, R. M., and Ferrara, A., 1999, Modeling time perception in rats: evidence for catastrophic interference in animal learning. In *Proceedings of the 21st Annual Conference of the Cognitive Sciences Society* (Hillsdale NJ: L. Erlbaum), pp. 173–178.

Glaze, J. A., 1928, The association value of non-sense syllables. *Journal of Genetic Psychology*, **35**: 255–269.

Hampshire, J. B., and Waibel, A., 1992, The meta-pi network: building distributed knowledge representations for robust multisource pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **14**: (7): 751–769.

Hintzman, D. L., 1984, MINERVA 2: a simulation model of human memory. *Behavior Research Methods, Instruments, and Computers*, **16** (2): 96–101.

Jacobs, R. A., Jordan, M. I., and Barto, A. G., 1991, Task decomposition through competition in a modular connectionist architecture: the what and where vision tasks. *Cognitive Science*, **15**: 219–250.

Jones, M. R., 1974, Cognitive representations of serial patterns. In B. Kantrowitz (ed.) *Human Information Processing: Tutorials in Performance and Cognition* (Hillsdale NJ: Erlbaum), pp. 187–229.

Jordan, M. I., and Jacobs, R. A. 1994, Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, **6**: 181–214.

Kruschke, J. K. 1992, ALCOVE: an exemplar-based model of category learning. *Psychological Review*, **99** (1): 22–44.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C., 1995, Why there are complimentary learning systems in the hippocampus and neocortex: insights from successes and failures of connectionist models of learning and memory. *Psychological Review*, **102** (3): 419–457.

McCloskey, M., and Cohen, N. J., 1989, Catastrophic interference in connectionist networks: the sequential learning problem. In H. G. Bower (ed.) *The Psychology of Learning and Motivation, Vol. 24* (New York: Academic Press), pp. 109–165.

McRae, K., and Hetherington, P. A., 1993, Catastrophic interference is eliminated in pretrained networks. In *Proceedings of the 15th Annual Conference Sciences Society* (Hillsdale NJ: L. Erlbaum), pp. 723–728.

Miller, G. A., 1956, The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, **63** (2): 81–97.

Mishkin, M., Ungerleider, L. G., and Macko, K. A., 1983, Object vision and spatial vision: two cortical pathways. *Trends In Neurosciences*, **6**: 414–417.

Poggio, T., and Girosi, F., 1990, Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, **247**: 978–982.

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

Restle, F., and Brown, E. R., 1970, Serial pattern learning. *Journal of Experimental Psychology*, **83**: 120–125.

Robins, A., 1995, Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, **7** (2): 123–146.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J., 1986, Learning internal representations by error propagating. In D. E. Rumelhart and J. L. McClelland (eds) *Parallel Distributed Processing: Explorations Microstructure of Cognition: Vol. 1. Foundations* (Cambridge MA: MIT Press), pp. 318–362.

Seidenberg, M. S., and McClelland, J. L., 1989, A distributed developmental model of word recognition and naming. *Psychological Review*, **96** (4): 523–568.

Shadmehr, R., and Brashers-Krug, T., 1997, Functional stages in the formation of human long-term motor memory. *The Journal of Neuroscience*, **17** (1): 409–419.

Sharkey, N. E., and Sharkey, A. J. C., 1995, An analysis of catastrophic interference. *Connection Science*, **7** (3–4): 301–329.

Shepard, R. N., 1987, Toward a universal law of generalization for psychological science. *Science*, **237**: 1317–1323.

Simon, H. A., and Kotovsky, K., 1963, Human acquisition of concepts for sequential patterns. *Psychological Review*, **70**: 534–546.

Appendix: Experiment stimuli

Control list 1:

ZOS	CEJ
CIJ	ZAH
KEF	TUZ
VAZ	VOZ
VEC	FOY
GOK	KEB
DIJ	GAK
HEG	JEH
ZEG	JIH
WUB	ZOT

Control list 2:

VEK	FEP
CEV	KIV
BIP	ZAT
NIZ	JID
GIK	KIZ
GEZ	KEZ
ZIN	GEK
GUK	JEC
TUV	VUT
PEF	JEZ

Inversion:

GOW	WOG
JID	DIJ
NIZ	ZIN
KUG	GUK
TUV	VUT
VUK	KUV
YOF	FOY
FEP	PEF
GEZ	ZEG
CEV	VEC

Vowel switch:

JIH	JEH
ZAT	ZOT
ZOS	ZAS
FEH	FIH
ZOH	ZAH
GEK	GIK
GOK	GAK
VAZ	VOZ
CIJ	CEJ
KEZ	KIZ