Episodic Memory in Connectionist Networks

Chris A. Kortge
Dept. of Psychology, Stanford University

Abstract

A major criticism of backprop-based connectionist models (CMs) has been that they exhibit "catastrophic interference", when trained in a sequential fashion without repetition of groups of items; in terms of memory, such CMs seem incapable of remembering individual episodes. This paper shows that catastrophic interference is not inherent in the architecture of these CMs, and may be avoided once an adequate training rule is employed. Such a rule is introduced herein, and is used in a memory modeling network. The architecture used is a standard, nonlinear, multilayer network, thus showing that the known advantages of such powerful architectures need not be sacrificed. Simulation data are presented, showing not only that the model shows much less interference than its backprop counterpart, but also that it naturally models episodic memory tasks such as frequency discrimination.

Introduction

One of the most obvious areas of application for connectionism is in modeling memory. McClelland & Rumelhart (1986) and others (e.g. Anderson, 1972) have shown how important properties of human memory--such as content addressability and prototype extraction--fall naturally out of parallel distributed models. Furthermore, such models have the potential of ultimately linking biological and psychological accounts of memory. However, as pointed out by Grossberg (1987) and McCloskey & Cohen (1989), one of the most powerful and popular classes of CMs--multilayer networks coupled with the backpropagation learning rule, or BPCMs for short--has had difficulty accounting for a fundamental aspect of human memory: the ability to learn and remember based on a single trial.

While BPCMs can learn a single pattern to a high degree of precision, they have seemed unable to do this for a *series* of such patterns without "forgetting" all but the last few. People certainly show retroactive interference to some degree, but as they can clearly remember more than the last handful of items from a long list, they don't exhibit the kind of catastrophic interference which seems, *prima facie*, to be inherent in BPCMs. Single-trial learning is central to many memory tasks--including all those based on a single presentation of a list, which is perhaps the most common experimental procedure--and thus no existing BPCM is a serious alternative to currently popular "global" theories of memory such as SAM (Raaijmakers & Shiffrin, 1981), and Minerva 2 (Hintzman, 1986). Ratcliff (1990) argues that serious problems exist for BPCMs in modeling recognition memory, in particular.

In this paper I propose a connectionist model of memory which can learn sequentially without catastrophic interference. I first describe the model, which is much like existing BPCMs except for its unique learning rule. Next I present simulation data showing that the model has no difficulty modeling memory tasks which are clearly episodic in nature. Finally, I use these data to argue that multilayer networks are still a viable approach to devising global theories of memory and learning.

The Model

The network architecture used by the model is a two-weight-layer "encoder" network (Ackley, Hinton, & Sejnowski, 1985), which maps an input vector of N elements to an output vector, also of N elements. Input patterns are vectors with elements of either +0.5 or -0.5, which form the activation values for the input units. Activations for other units are computed using a symmetric logistic function. Thus

$$a_j = [\, 1 / (1 + \exp(-\textstyle\sum_i a_i w_{ji})) \,] - .5$$

where a is the activation value (j indexes the current layer, i the previous layer), and w_{ji} the weight on the connection from unit i to unit j. Unlike a typical network of this type, no biases are used in computing activations.

The network's task is to learn to reproduce each presented input pattern at the output, given fewer hidden units than N. In the simulations reported here, performance after learning is measured by the "match" between the input and output vectors, where "match" is defined as the dot product between input and output, divided by N, and multiplied by 4 to give a number between -1 and 1. Match values may then be compared for different classes of items, such as those items appearing versus not appearing in a studied list. The use of such an architecture in memory modeling is not new; Ratcliff (1990) uses a very similar one in his diagnosis of the interference problem. The main contribution of the present paper centers on the new algorithm used to train the network, which is now described.

Two key points need to be made in describing the learning rule used. The first point is that, in associative network learning in general, orthogonality of inputs is sufficient for eliminating interference. This means that if pattern A is trained to give an output A*, subsequently training pattern B to produce *any other output* will not change the A-to-A* mapping as

long as B is orthogonal to A. This is true in most one-weight-layer networks, and as elaborated below, can be made true to a large degree in multi-layer nets as well, by careful selection of parameters. Whether it is reasonable to assume that a model's inputs will all be orthogonal to one another, though, is another question, and is addressed in the Discussion section below.

The second key point is that gradient descent (see Rumelhart, Hinton, & Williams, 1986) is not necessary for reducing the error on a given pattern. If we look at learning as moving through the space of possible weight combinations, only one direction corresponds to doing gradient (steepest) descent. On the other hand, *half of all possible* directions will still reduce the error, although not as fast as following the gradient would. This idea is quantifiable--the dot product of the actual change in the weights with the gradient indicates the speed with which the error is reduced. The important thing here, though, is that by relaxing the constraint that the gradient be followed (that error be reduced as fast as possible), and instead only requiring that the error be reduced to *some* degree, we gain "room" for adding another constraint to our network's learning process--namely, one which reduces interference.

The standard backpropagation learning rule, which uses gradient descent, operates essentially by associating activated units in one layer with desired outputs at the next layer (to the degree the current outputs are wrong). The idea behind the present algorithm, on the other hand, is roughly to use a *subset* of the activated units in the association. In particular, only the "novel" activations are used, as defined below. Conceptually, the idea is this: when the network makes an error, we would like to blame just those active units which were "responsible" for the error--blaming any others leads to excess interference with other patterns' outputs. And, the argument goes, because "familiar" things are well learned (by definition, in a memory system) we might reduce interference with well-learned information by assuming that the novel aspects of the input are "responsible" for the output error; that is, by changing weights only from them, and not from the familiar parts.

The precise definition of "novelty" is crucial, of course. The basic approach, though, is to substitute a "novelty vector" for the actual activations during standard backprop learning (not during feedforward processing). Thus the change to a particular weight, w_{ji}, is now

$$\Delta w_{ji} = \eta \, \tilde{a}_i \, \delta_j ,$$

where $\tilde{a}_i$ is the "novelty" of the activation of unit i, η is the standard learning rate parameter, and δ_j is the delta signal as computed by standard backpropagation. This will allow both of our desired constraints to be met: First, the novelty vectors (one for the input layer, and one for the hidden layer) are chosen so that each tends to be orthogonal to highly familiar activation patterns at that layer. This property will allow reduced interference when the weight changes are made, by insuring that weights from well-learned patterns are changed less. Second, a novelty vector is generally a "part" of the actual activation vector it replaces, in some sense. As elaborated below, this insures that when we create associations by changing the weights just from the novel parts, we are still reducing the error on the pattern at hand (although not as fast as with gradient descent).

The novelty vectors used were as follows: The novelty vector for the input layer is simply the input (target) vector minus the output vector. This makes intuitive sense: because the task of the network is to reproduce previously seen inputs at the output, the output error should provide a rough indication of which aspects of the input are novel. As for the hidden layer, its novelty vector is obtained by feeding the input novelty vector through the first layer of connections (just as if it were another input pattern), and using the resulting hidden unit activations as the hidden novelty vector. (As explained below, the hidden novelty vector is taken to be zero in certain cases.) The justification for using this particular vector is not strong; it is simply that this hidden novelty should tend to correspond to hidden activation patterns which have not previously occurred (because the input novelty which produced it is *itself* a pattern which isn't highly learned).

These particular definitions of "novelty" are not necessarily optimal; they are only heuristics. They do possess the characteristics described above, though. First, they tend to be orthogonal to highly familiar activation patterns at the corresponding layer, even when the current activation pattern is *not* orthogonal to previous ones. This is because (1) output errors tend to be orthogonal to well-learned outputs, and (2) since an encoder network is being used, and thus there is a correspondence between inputs and outputs, we can readily define input novelty in terms of output error.

Second, these novelty vectors satisfy the constraint that the new algorithm still reduce the output error. It can be shown that substituting *any* vectors for the activation vectors during backpropagation learning will still allow the error (which is computed using the *true* activations) to be reduced, *if* (but not only if) each activation vector **a** is replaced by a vector *correlated* with **a**. It is readily observed that when using binary +.5/-.5 targets, as here, the sign of *target* − *output*, the input novelty, is always the same as the sign of *target*, for each unit. Thus the input activations and the input novelties are correlated. As for the hidden novelty vector, it also tends to be correlated with the actual hidden activations.

However, this is not guaranteed, so the hidden novelties are set to zeros when this is not true (an infrequent occurrence), thus restricting learning to the first weight layer.

In sum, then, this new learning algorithm is much like BP, but with a more focused assignment of blame. By blaming novel things more than familiar things when an error occurs--where "novelty" is determined by the evolving network memory itself, in a context-dependent manner--the network can modify to a greater degree those connections which don't much affect the storage of well-learned patterns. As a result, storage of a new pattern requires relatively little disruption of existing knowledge. Because the gradient is not followed, this advantage is obtained at a cost of lower learning speed. However, "learning speed" in this case means *within a pattern presentation*, which in human terms might correspond to the amount of time a stimulus is viewed by a subject. But since we have no idea how much network learning (e.g. magnitude of weight change) should correspond to a given stimulus duration in a human learning experiment anyway, this is not a problem for the model. What is important is that the present model exhibits improved learning *across* patterns, in that old patterns need not be represented many times to allow remembering of a long list. This is shown next.

Simulation Data

Forgetting functions

Perhaps the simplest way to assess interference is to present a list of patterns, training each pattern in turn by some amount, and, after one run through the list, to examine performance as a function of serial position of the list items. This is the basic procedure used in many experiments on human memory, and is also the procedure Ratcliff (1990) argues must be used in modeling such experiments (i.e., if a list is not repeated for a person, it should not be repeated for our network model). The present model was tested against standard BP on such a task, using identical network architectures. These networks were as described above, using 32 input, 16 hidden, and 32 output units (a "32-16-32" net). For each simulated subject, 16 binary patterns were constructed--random vectors, with elements +.5 with probability 1/2, otherwise -.5.

For each run, 15 of the 16 patterns were actually learned. The learning rate, η, was .5, and the initial weights of the network were normally distributed with mean zero and variance (σ^2) = .25. Each pattern in turn was repeatedly presented--meaning simply that multiple learning steps were taken on each--until either the mean squared error over the output units reached .01, or a pre-set maximum of 1000 steps was reached. (From now on, each set of steps on a single pattern will be considered one "presentation".) After the 15 pattern presentations, the match value between input and output was tested for each pattern, including the unpresented one. There were 250 simulated Ss in each condition. Figure 1 shows the resulting serial position curves for the novelty rule, backprop, and another instantiation of backprop as tested by Ratcliff (1990), which only studied eight single items.

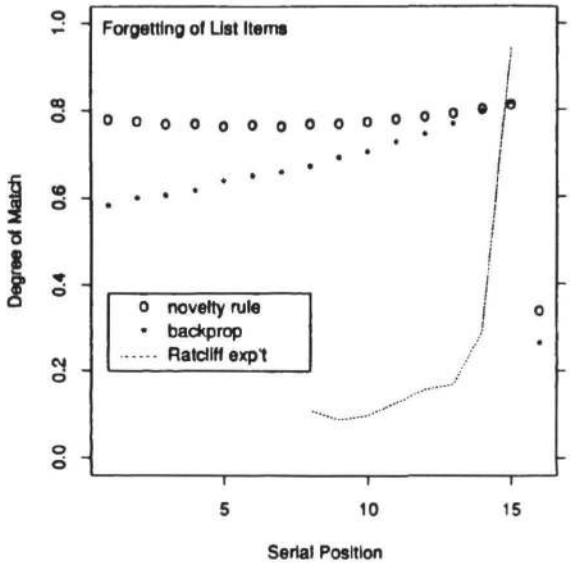

Figure 1. Degree of input/output match as a function of serial position, after one list presentation. Position 16 represents the unpresented item. The Ratcliff data begins at position 8 for comparison purposes.

The first thing to note in the figure is that there is much more forgetting in the Ratcliff experiment than in either of the others. Two aspects of Ratcliff's network probably contributed to its worse performance: First, asymmetric (0 to 1) activation functions were used, which induces a larger average correlation between activation vectors. This generally leads to higher interference, as already noted. Second, smaller starting weights were used (uniform +.3 to -.3). David Rumelhart (personal communication) has suggested that larger initial weights tend to lead to lower interference, because training tends to move activations further into the tails of the sigmoid function. Because the derivative of the sigmoid is near zero there, subsequent changes of weights to the "saturated" units will be small, meaning that interference with the patterns represented by those units will be less. This analysis was supported by comparing separate runs of backprop on the present experiment using σ^2 values of .01, .25, and 1.0.

It is also apparent from the figure that forgetting is even less when using the novelty-based learning rule.[1] This result must be qualified, though. Further

[1] In addition, there appears to be a slight primacy effect with the

simulations showed that both backprop and the novelty rule improved significantly on this task with increasing network size. It may well be, then, that with a large enough net there would be virtually no interference with *either* algorithm on this task. Furthermore, importantly, the question of what size net is "proper" for modeling memory remains open, as noted in the Discussion section below.

However, while backprop improved greatly with larger nets, the patterns used were also uncorrelated on average, which may have helped. The next experiments used patterns which were distortions of a single (random) prototype pattern, and thus were explicitly correlated. Each distortion was generated by flipping any given bit of the prototype with probability .25. Figure 2 compares the serial position functions obtained with such patterns for two sizes of backprop networks (32-16-32 and 128-64-128), and the 32-16-32 novelty network used above. Other parameters were as above.

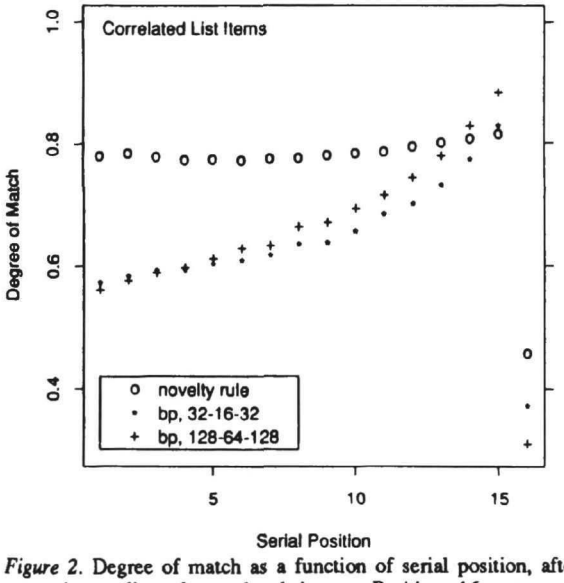

Figure 2. Degree of match as a function of serial position, after presenting a list of correlated items. Position 16 represents untrained items which were correlated with the presented items.

Comparing the two backprop nets, while the large-network curve ends up higher (possibly due to a tendency to produce more learning per step), its slope is also steeper; thus overall, forgetting is more pronounced if anything for the larger net (note that new-item match is lower for the larger net, though). A 256-128-256 net, and nets with larger σ^2, gave similar results. These results were perhaps predictable: While forcing activations further into the sigmoid tails can reduce interference--by making delta vectors more orthogonal to learned outputs--this should not reduce any interference caused by correlated *inputs*. There is no apparent way to remove correlations among the inputs simply by changing the network configuration.

The novelty-rule network, on the other hand, exhibited very little interference even with these correlated items. Moreover, further runs showed the novelty rule to *improve* with network size: absolute match values were approximately unchanged, but new-item match values decreased significantly, so that forgetting in terms of discrimination was less. Thus it appears that even with correlated patterns, multilayer network models need not suffer from extreme interference.

Varying amount of learning

In the next experiment, discrimination between "old" (studied) and "new" (unstudied) items was examined as a function of amount of learning, as measured by number of learning steps taken, or the "duration" of each pattern presentation. This is important to check, because while human data shows increasing discrimination with increasing item learning time, such an improvement is not predicted by some current memory models. Simple linear models, for example, predict that the variance in output to new items increases with the mean old-item output as learning amount increases, so as to keep d' constant (Shiffrin, Ratcliff, & Clark, 1990). Also, Ratcliff (1990) has shown that several variants of the backprop-based encoder network do not predict a strictly increasing d' (the actual patterns are complex, often nonmonotonic, and dependent on the parameters used).

A 32-16-32 architecture was used as in the forgetting simulations. Other parameters were $\sigma^2 = .25$, and $\eta = .1$. However, rather than training each list item to a criterion as before, each item was simply given a fixed number of learning steps before proceeding to the next item. This number of steps was varied (between lists) from 1 to 512, by powers of 2. Two types of patterns were tested, for both standard backprop and the novelty variant: "uncorrelated", with elements +.5 and -.5 equally likely; and "correlated", where list items were distortions of a single prototype, as described above. Within each learning amount, 100 lists of 16 items were trained for each condition. Performance was compared on the learned items and 16 new items using d' (mean old-item match minus mean new-item match, divided by standard deviation of new-item match), where new items were generated the same way as list items (from the same prototype, in the "correlated" conditions).

As depicted in Figure 3, d's were increasing for the most part. (The very slight decrease over the last interval in the novelty/correlated condition was insignificant.) Furthermore, in a replication of all

novelty rule, with items from the beginning of the list giving better matches than those in the middle. More work is necessary to determine the significance of this.

conditions, using $\sigma^2 = 1.0$, there were *no* decreases, and a similar overall pattern was obtained. Note that in all conditions the increases extended into or past the range of d's typical in human experiments (about 1 to 3). Mean match values after training the 512-step lists ranged from about .70 to .78. Backprop was somewhat worse overall than the novelty rule, presumably due to interference effects. In sum, Ratcliff's conclusion that increasing d' is a problem for the encoder network must be qualified. In particular, further simulations suggested that initial weight variance is critical: using $\sigma^2 = .01$ rather than .25, nonmonotonic functions much like Ratcliff's were obtained with backprop (the novelty-rule function was still monotonic). There is a degree of this nonmonotonicity apparent in the bp/uncorrelated condition in the graph, as well.

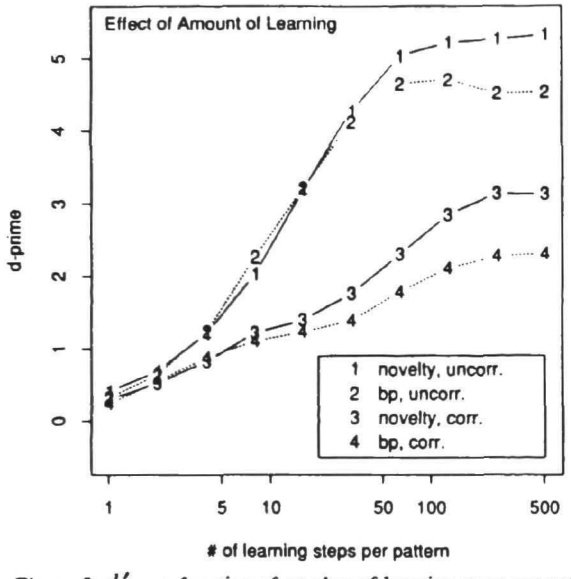

Figure 3. d' as a function of number of learning steps per pattern in a once-presented list. Note the logarithmic scale.

Frequency discrimination

After being presented with a list of items, some of which are repeated various numbers of times (at various intervals), people can give fairly good estimates of the frequencies with which items were presented. This task can be viewed as a generalization of the recognition task, wherein the possible frequencies are limited to zero and one. Hintzman (1988) has called frequency judgment "a quintessentially episodic memory task", presumably because information concerning individual item presentations (namely their occurrence) is required in order to respond correctly.

The present novelty-based model was applied to this task, using a paradigm similar to one used by Hintzman (1988) in testing his multiple-trace memory model, Minerva 2, on frequency discrimination. A size 48-9-48 network was used, with $\eta = .5$ an $\sigma^2 = .25$. Twenty-four random patterns were generated, with four assigned to each of six frequency conditions, 0 to 5. Each pattern was copied the proper number of times, and the resulting list of 60 items was randomly permuted. Items were then trained one-by-one, with only a single learning step taken on each (thus considerably less training was done per pattern than in the forgetting experiments above). After training, match values were tested for all 24 patterns. This procedure was replicated for 5000 simulated Ss, resulting in distributions of match values for each of the six frequencies. From these distributions forced-choice data were computed, giving the proportion of errors the network would make if required to pick the more (or less) frequent item from a pair taken from the original 24. These error data are shown in Figure 4.

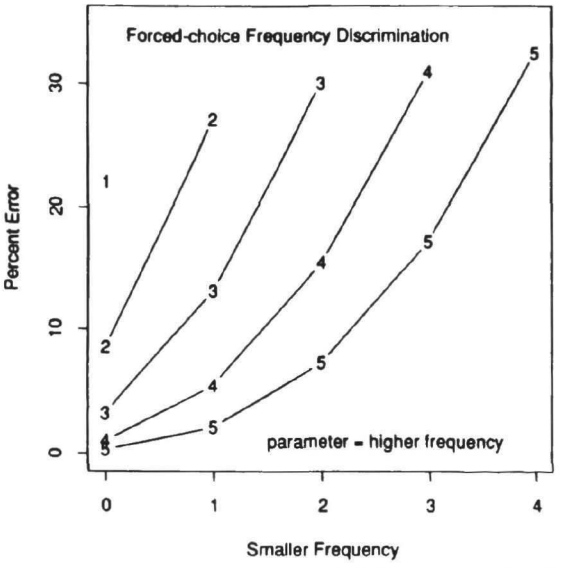

Figure 4. Errors in simulated forced-choice data, as a function of the frequency of the less frequent item in a test pair. Data were generated from distributions of match values corresponding to each frequency.

Overall the model predicts that (given appropriate response criteria) discrimination between two frequencies improves with greater frequency difference, and holding frequency difference constant, discrimination worsens with increasing frequencies. This same pattern holds with people (Hintzman & Gold, 1983), when results for the "choose the more frequent" and "choose the less frequent" instructions are collapsed. Interestingly, the results obtained here are very nearly indistinguishable from those obtained with Minerva 2; in fact, although each matches the human data well, the models match each other much better than either matches the human data. For this task, then, there seems to be no theoretical advantage to maintaining a distinct memory trace for each individual "event" to be remembered. Clearly, the present connectionist model does as well by "strengthening" each repeated item (by increasing the relevant weights). (Backprop was also able to model this particular task.)

List-specific frequency judgment

One might rightfully argue that there is more to making frequency judgments than simply assessing the familiarity of an item, though. In particular, human subjects are faced with the more difficult task of not only having to discriminate different frequencies that occurred *within* an experimental list, but *also* having to discriminate the list presentations from all the times the same items have been observed *outside* the experiment. Thus, for example, people are quite good at discriminating rare words which appeared in a list from common words which did not appear, even though presumably the "baseline" familiarity is much higher for the common words. Another example is provided by an experiment by Hintzman & Block (1971), in which people were asked to make *list-specific* frequency judgments. Even though two different lists were made up of the same items, subjects could give good estimates of an item's frequency in one list almost independently of its frequency in the other list. In a sense, this task is even more "quintessentially episodic" than standard frequency judgment; not only must information about number of occurrences be retained, but this information must be distinguishable on the basis of instructions to limit the relevant context.

The present model was tested on this task using a 120-30-120 net, with $\eta = .3$ and $\sigma^2 = .25$. An approach much like Hintzman's (1988) was used. In order to model different lists, 80 input units represented list context, and each list was assigned its own random context vector which was presented along with every item in that list. Twenty-seven random 40-element vectors were generated, and three of these assigned to each of the 9 possible combinations of the frequencies 0, 2, and 5 across the two lists. Thus each of the three "2-0" patterns appeared twice in the first list and zero times in the second list, and similarly for 0-0, 0-2, 0-5, 2-5, and so on. Once the proper number of copies were made, each list was randomly permuted and then trained, with one learning step per item.

Figure 5 shows the match values obtained after training both lists, averaged over 500 simulated Ss, along with the data from H & B linearly transformed to match the model's output values (using a least squares fit over the 9 means). Each pattern was tested separately with each context vector, and data for the two lists was collapsed.[2] Thus for example the "target-frequency = 5, nontarget-frequency = 2" data point represents match values obtained by testing the 5-2 patterns in context 1, and the 2-5 patterns in context 2.

[2] The model showed a reliable primacy effect, with higher matches overall on the first list, but it was not large and did not affect the conclusions.

In the graph, perfectly flat (separated) lines would mean that discrimination was perfect; i.e., that there was no interference from the nontarget list in making frequency judgments. Hintzman (1988) proposes measuring performance on this task by a "discrimination index" (DI), obtained by dividing the variance among means accounted for by nontarget frequency by that accounted for by target frequency. Such a measure could vary from 0 (perfect list discrimination) to 1 (no discrimination). H & B's subjects showed a DI of .097, and the present simulation gave a DI of .093. Correlation of the model with the H & B means was .998. (A total of about five simulations of this experiment preceded this particular one, so not a lot of fitting was involved.)

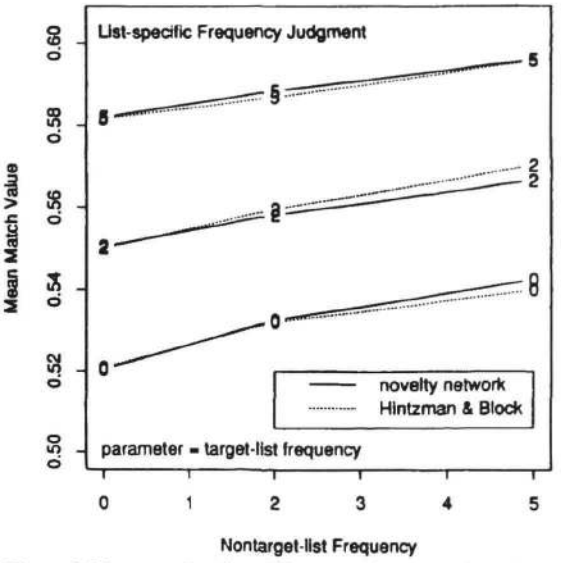

Figure 5. Mean match values ("frequency judgments") for items in a particular list, as a function of frequency of the same item in the other list. Hintzman & Block's subjects' data was linearly scaled to match the network's output values.

The present model had no trouble, then, even with this highly episodic memory task. (Backprop also had no trouble, except when interference effects were large due to use of correlated items.) The novelty-based network did require many more context elements than did Minerva 2, however, in order to match human performance (as with Minerva 2, discrimination improved with additional context elements). This might seem disadvantageous, but it isn't clear how much of a person's internal representation of an episode is devoted to "context", anyway. Moreover, Minerva 2 does not attempt to model the existence of extra-experimental memories. It seems reasonable that the more occurrences of an item are stored in memory, the larger must be the number of possible contexts in order that any small subset of those memories be accessed. Put differently, the more background noise, the harder the discrimination task becomes. Thus a version of Minerva 2 which modeled prefamiliarization might require more

context elements as well. The present model, on the other hand, does all its learning in terms of what is already known; i.e., all previous knowledge is contained in the model, although here this pre-existing memory was modeled simply with random weights. Thus it has more of a burden in trying to discriminate the experimental presentations from background "memories".

Discussion

As a whole, the results presented here argue strongly that episodic memory tasks are not beyond the modeling capabilities of multilayer networks. First, it was shown that the catastrophic interference exhibited by many of these networks is readily avoided. Three different tricks were used to accomplish this: (1) increasing the variance of the starting weights, which, by making more use of the nonlinearity of the activation functions, tends to protect from change those weights leading to units being "used" to represent old items; (2) increasing network size, which provides more units to *be* "used" in representing old items, and also increases the variance of initial activations (by increasing the fan-in to the units) which should have an effect similar to increasing the initial weights; and (3) using a new learning rule, which increases orthogonality for learning purposes while still reducing error. While use of the first two tricks suggested that arbitrarily low interference might be obtained with backprop using uncorrelated (on average) patterns, only by using the new learning rule did this seem possible for explicitly correlated patterns.

Second, it was shown that the increasing d' exhibited by people with increased learning per item can be modeled by backprop, within certain parameters, and by the novelty-based learning rule, with no observed need for limiting parameters. This attenuates the conclusion made by Ratcliff (1990) that this is a problem for backprop networks. (Ratcliff does note the qualitative differences changes in parameters can make.)

Third, the novelty-based network (and backprop to some extent) was shown capable of modeling a simple frequency discrimination task, and a more difficult list-specific frequency judgment task. While, unlike recall tasks, these tasks are based only on scalar "familiarity value" outputs, they still depend on the ability to maintain information about individual episodes, and thus seem like good indicators of episodic memory.

Disadvantages

Because of the preliminary nature of the present model, it is hard to say much about its explanatory disadvantages. It is reasonable to ask at least, though, whether the tricks used herein are justified from a theoretical standpoint. Increasing initial weight variance doesn't seem problematic: people clearly bring a lot of knowledge to an experiment (whether random weights capture this knowledge is debatable, of course). Increasing network size seems fairly reasonable as well; obviously the number of a person's neurons dwarfs the number of items in any memory experiment. However, it must be noted that, performance level held equal, smaller networks seem to generalize better to new experience than large ones (consistent with general overfitting arguments). Note, though, that "interference" and "generalization" are in essence two sides of the same coin. Each implies learning of one pattern affecting the output of another; the difference is whether such effects are "good" or "bad", given the task at hand. It is an open question, then, to what extent future CMs will exhibit high "good generalization" and low "bad generalization".

As for the novelty-based learning rule, no obvious explanatory disadvantages (relative to other CMs) have surfaced yet. Indeed, it seems clear that people *do* tend to focus, in some sense, on unusual aspects of their environment. Furthermore, it seems like successive inputs to a human memory system would tend to be highly correlated, if anything, based on continuity of the environment. Thus the novelty rule (or something similar) may well be necessary for modeling human memory in multilayer nets without high interference. On the negative side, preliminary runs suggest that the novelty rule is not as good as BP at storing information in small (relative to the task) networks, when very large amounts of learning per pattern are used. This need not be problematic for modeling, though, since as noted there is no pressing theoretical limitation on a modeling network's size.

Relation to other models

Several points need to be made regarding previous models. First, a large number of connectionist (or connectionist-flavored) models do not suffer from the catastrophic interference problem as do BPCMs. These include holographic models such as CHARM (e.g. Eich, 1982), and the various ART models (e.g. Grossberg, 1987), among others. The present research speaks to these models only indirectly, insofar as it lends support to a class of potential competitors (multilayer error-reduction networks). It is also important to note that the basic idea behind the learning rule introduced here--that focusing on novelty in learning can reduce interference--is not new. Grossberg (1987) has argued that this is necessary to allow stability of learning in a changing environment. Otwell (1990) has used substitution of novelty vectors (of a different sort than here) in backprop learning. And in an intriguing parallel, Holyoak, Koh, & Nisbett (1989) use an "unusualness heuristic" in generating "exception rules" in their rule-based theory of animal conditioning, which

function to "censor useful but imperfect default rules, protecting them from loss of strength". This seems roughly like another way to say "reduce interference". The present paper, then, might be used to argue that the difference between rule-based and connectionist accounts of learning is not so clear as Holyoak et al. suggest.

There are many similarities between the present model and Hintzman's multiple-trace model, Minerva 2 (1988). For instance, inputs are represented as feature vectors in each, and each allows for two kinds of outputs: a scalar familiarity value, and a vector representing a "retrieved memory". These similarities have led to the frequency judgment experiments herein being modeled much as in Minerva 2. The models' storage assumptions, however, are quite different: while Minerva 2 maintains distinct memory vectors corresponding to each experienced "event", the present model, as other CMs, superimposes all memories on a single set of connection weights. Thus it is striking that the agreement in their predictions noted above is as strong as it is.

Hintzman allows for the possibility that associative matrix models might account for the range of frequency-judgment data exhibited by people, but opts for a multiple-trace view for various reasons--such as the accessibility of individual event information, and the ability to activate individual traces as a nonlinear function of their similarity to a probe (Hintzman, 1988; 1986). The present research argues, though, that given some distinguishing context cues, individual event memories can also be accessed by a network. Also, because of the nonlinear activation functions of the present model, it too has the potential to exhibit nonlinear generalization as people do (pilot work on prototype extraction experiments has supported this claim). Thus there is no *obvious* reason why multiple-trace theories such as Minerva 2 should possess any inherent explanatory advantage over CMs.

Conclusion

It has been argued herein that a potentially major problem with multilayer network models--that they exhibit catastrophic interference--is in fact not inherent in these models at all. By using the learning rule introduced here, and in certain cases using even simpler devices, these models can learn individual associations to a large degree with little disruption of prior knowledge. Given this, there is no obvious reason why an encoder network-based alternative to current global models of memory could not evolve. Of course, the present paper has barely scratched the surface of such an undertaking, and there is much more to be said, surely both pro and con, about such memory models. Nonetheless, a non-obvious capability of these networks--the ability to learn, remember, and later make use of information about individual event information--has been shown to exist. Taken together with previous connectionist work, and the tantalizing prospect of meshing biological and psychological accounts of memory, this suggests strongly that connectionism has a solid place in future memory research.

Acknowledgments

I wish to thank David Rumelhart for his ongoing help, as well as helpful input on the present paper, and David Bryant and Gary Cottrell for their extremely helpful reviews. I also thank Roger Ratcliff for providing me with a draft of his paper, and apologize in advance for any confusion arising from my use of this draft rather than the published version.

This research was partly supported by a National Science Foundation Graduate Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9*, 147-169.

Anderson, J. A. (1972). A simple neural network generating an interactive memory. *Mathematical Biosciences, 14*, 197-220.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review, 89*, 627-661.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science, 11*, 23-63.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411-428.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*, 528-551.

Hintzman, D. L. & Block, R. A. (1971). Repetition and memory: Evidence for a multiple-trace hypothesis. *Journal of Experimental Psychology, 88*, 297-306.

Hintzman, D. L. & Gold, E. (1983). A congruity effect in the discrimination of presentation frequencies: Some data and a model. *Bulletin of the Psychonomic Society, 21*, 11-14.

Holyoak, K. J., Koh, K., & Nisbett, R. E. (1989). A theory of conditioning: Inductive learning within rule-based default hierarchies. *Psychological Review, 96*, 315-340.

McClelland, J. L. & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*, 159-188.

Otwell, K. (1990). Accelerating backpropagation learning with novelty-based orthogonalization. *Proceedings of the International Joint Conference on Neural Networks*, Jan. 1990, Vol. 1. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Raaijmakers, J. G. W. & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*, 93-134.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions (DRAFT). To appear in *Psychological Review, 97*, March 1990.

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, *Parallel Distributed Processing*, vol. 1. Cambridge: MIT Press.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 179-195.