(To appear in the *Proceedings of the 16th Annual Cognitive Science Society Conference)*

Dynamically constraining connectionist networks to produce distributed, orthogonal representations to reduce catastrophic interference

Robert M. French
Willamette University, Salem, OR 97301
french@willamette.edu

Abstract

It is well known that when a connectionist network is trained on one set of patterns and then attempts to add new patterns to its repertoire, catastrophic interference may result. The use of sparse, orthogonal hidden-layer representations has been shown to reduce catastrophic interference. The author demonstrates that the use of sparse representations may, in certain cases, actually result in worse performance on catastrophic interference. This paper argues for the necessity of maintaining hidden-layer representations that are *both* as highly distributed and as highly orthogonal as possible. The author presents a learning algorithm, called context-biasing, that dynamically solves the problem of constraining hidden-layer representations to simultaneously produce good orthogonality and distributedness. On the data tested for this study, context-biasing is shown to reduce catastrophic interference by more than 50% compared to standard backpropagation. In particular, this technique succeeds in reducing catastrophic interference on data where sparse, orthogonal distributions failed to produce any improvement.

Introduction

It is well known that when a connectionist network is trained on one set of patterns and then attempts to add new patterns to its repertoire, catastrophic interference — in other words, the complete loss of all of its previously learned information — may result (Ratcliff, 1989; McCloskey & Cohen, 1990; Hetherington & Seidenberg, 1989). This type of radical forgetting is not only psychologically implausible — learning the names of three new people does not cause us to forget the names of everyone else we have ever met — but also poses significant problems for applications-oriented uses of connectionist networks.

French (1991) suggested that catastophic forgetting was caused by overlapping (i.e., non-orthogonal) patterns of activation at the hidden layer. A learning technique, called activation sharpening, was proposed that reduced this overlap by creating "sparse" representations (i.e., representations consisting of a few highly active nodes and many nodes with activation levels close to 0). This technique did produce a significant decrease in catastrophic forgetting. A number of authors (Murre, 1992; McRae & Hetherington, 1993) also developed techniques that resulted in improved orthogonalization of representations at the hidden layer and they, too, observed similar decreases in catastrophic interference. Lewandowsky & Goebel (1991) developed a technique for orthogonalization of input vectors that also resulted in improved orthogonality of representations at the hidden layer. All of these techniques have been shown to decrease catastrophic interference.

In this paper I will argue that increased orthogonality of hidden-layer representations is only part of the story. When attempts to produce representational orthogonality excessively restrict of the distributedness of representations, there may be no reduction (or even an increase) in catastrophic forgetting compared to standard backpropagation (Figure 3).

The necessity of orthogonality *and* distributedness

The central claim of this paper is that, in order to reduce catastrophic interference, learning algorithms for connectionist networks must dynamically produce hidden-layer representations that are, insofar as possible, both:

- highly orthogonal *and*
- highly distributed across the hidden-layer.

I will introduce a simple recurrent learning algorithm, called context biasing, that produces hidden-layer representations that satisfy both of these constraints. To test this algorithm, I selected a data set for which sparse representations, even though highly orthogonal, failed to reduce catastrophic forgetting (and in many cases aggravated it). Context biasing was found to reduce catastrophic forgetting for this data by more than 50% compared to standard backpropagation.

Orthogonalization using activation sharpening

Activation sharpening (French, 1991) reduced catastrophic interference by producing sparse hidden-layer representations. The technique is best explained by considering *one-node* activation sharpening. On each input/teacher presentation P, there is first a standard feedforward–backpropagation pass. Then the same input pattern is then fed forward from the input layer to the hidden layer, producing a particular activation pattern at the hidden layer — the "natural" hidden-unit representation. A "target" hidden-layer representation is then created by slightly increasing the activation of the single most active node and decreasing the activation of all of the other nodes. The difference between the "target" and the "natural" hidden-layer representations serves as an error signal analogous to the output/teacher error signal in standard backpropagation. This hidden-layer error is backpropagated from the hidden layer to the input layer, changing the input-to-hidden-layer weights appropriately. Under this "backpropagation-plus-halfpropagation" technique, the hidden unit representations are gradually "sharpened", with each representation consisting of one highly active node and the other nodes having very little activity. *k-node* sharpening involves modifying hidden-unit representations by increasing the activation of the *k* most active nodes (instead of only the single most active node in one-node sharpening), and decreasing the activation of the others.

French (1991), Murre (1992) and Lewandowsky & Goebel (1991) have shown that this technique was effective in reducing activation overlap among hidden-layer representations which, in turn, decreased catastrophic forgetting. Unfortunately, all of the databases used to test this particular orthogonalization technique were very small. Problems became apparent only when I attempted to use the technique on a significantly larger database.

Problems with sparse representations

The problem with orthogonalizing techniques that rely on sparse representations is that, while this may indeed produce the desired orthogonality among hidden-layer representations, they also reduce the effective dimensionality of the hidden-layer "representation space" (i.e., decrease the number of patterns that can be encoded by the hidden layer).

If a network has ten hidden units and one-node sharpening is implemented, for example, each hidden-unit representation will gradually be forced to have one unit with an activation close to 1 while the remaining nine units will have little or no activation. This means that, effectively, only ten distinct hidden-layer representations are possible. Therefore, if fifteen different categories of patterns are to be learned, the network will have great difficulty accomodating all of them in its hidden-layer representation space. This will mean that the network will invariably take a long time to converge, and probably will not converge at all. The unavoidable conclusion is that networks with sparse hidden-layer representations have a diminished capacity to categorize. This becomes a problem when the number of different classifications that the network must make exceeds the number of possible representations at the hidden layer.

Lewandowsky & Goebel (1991) and Lewandowsky & Shu-Chen (1993) have claimed that the ability of a network to generalize remains basically unaffected in networks that achieve representational orthogonality by activation sharpening (or any other technique that produced sparse representations consisting of a few highly active nodes and many completely inactive nodes). However, the situation turns out to be somewhat more problematic. Lewandowsky *et al.* showed that overall levels of activation on output are largely unaffected by the use of sparse, orthogonal hidden-layer representations and from this they concluded that the ability of the network to generalize would not be affected.

The difficulty is that, while sparse orthogonalization techniques may not change the *overall level* of activation, the *patterns* of output activation are radically modified. As the following examples show, the use of sparse orthogonal representations *does* result in information loss across the hidden layer and, ultimately, impairs the network's ability both to categorize and to discriminate among disparate input patterns.

Consider a **6-2-3** backpropagation network that must learn the following three associations:

110000 —> **100**;
001100 —> **010**;
000011 —> **001**.

The standard backpropagation network can easily learn these three pairs. But if, instead of allowing the activations in the hidden layer to range freely from 0 to 1, assume that only two hidden-layer representations are allowed — namely, **0 1** and **1 0**? (This would be the theoretical result of one-node activation sharpening in this case.) Now the network must learn all three associations with only two orthogonal hidden layer representations and it will fail to converge.

Now, let us consider the problem of pattern discrimination using a slightly different example. Assume we have an **8-3-5** network and we want it to learn three patterns:

11110000 —> **10000**;
00111100 —> **00100**;
00001111 —> **00001**.

The network then learns these patterns with orthogonalized hidden-layer vectors **1 0 0**, **0 1 0**, and **0 0 1**. This will cause three of the hidden-to-output weights to have high positive values, and all of the others to have high negative values (Fig. 1).

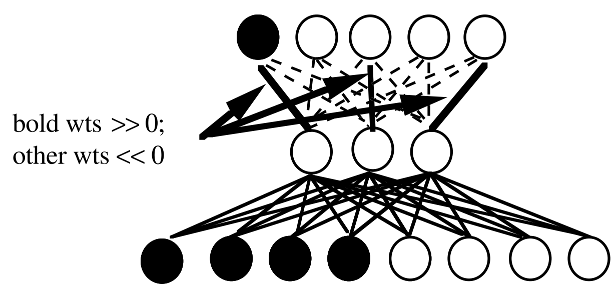

Figure 1. The effects on the hidden-to-output weights of one-node sharpening

Forcing the hidden-layer representations into the three all-or-nothing representations above has the effect of causing three of the hidden-to-output weights (those indicated in bold in Figure 1) to be very large, while all of the others will have large negative values. With these orthogonalized hidden-layer vectors, the second and fourth output nodes will effectively cease to participate in the output, since their activation levels will remain unchanged regardless of the input. This means that the network will produce "3-dimensional" output for all input patterns (Figure 2). A standard backpropagation network, on the other hand, will produce a much richer, 5-dimensional output response pattern for each of the new input vectors. Collapsing five output dimensions to three means that some of the patterns that could have been discriminated by the standard network will no longer be discriminated by the orthogonalized network. For this reason, it is necessary to relax the constraints on orthogonality. To produce output that is better than 3-dimensional, we *must* have hidden representations that are not always orthogonal. However, if we completely renounce attempts at orthogonality, severe foretting will result. An orthogonalization process that ensures that as large a number of nodes as possible will participate in the representations should have the desired effect of reducing catastrophic interference without completely crippling the network's ability to generalize, categorize, etc. But is it possible to achieve a high degree of both distributedness *and* separation of hidden-layer representations?

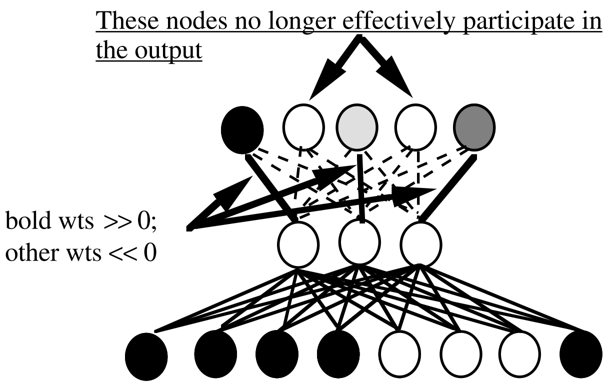

Figure 2. Restricting the dimensionality of hidden-layer representations decreases the dimensionality on the output layer

Testing sparse, orthogonal representations on a larger data set

In order to further study the use of sparse orthogonalized representations, I used data from the 1984 Congressional Voting Records (Murphy & Aha, 1992). The voting records of each member of Congress, along with his or her party affiliation, is given in this database. I trained a 16-10-1 feedforward backpropagation network to associate 50 different voting patterns with party affiliation. I then invented a small set of ten "maverick" members of Congress, members who, on six key issues, voted like Democrats but declared themselves to be Republicans or vice-versa. I had the network learn this new set and, as expected, when I tested its performance on the original set of fifty associations, the network had completely forgotten them. I then retested the network for savings. The speed with which the network relearned the original data was used to measure how completely the network had forgotten the original data.

When sparse representations were used, the network performed significantly worse than when it used more highly distributed representations (Figure 3.) The reason for this can be seen by looking at the average hidden-layer activation profiles for Democrats and Republicans produced by two-node sharpening (Figure 4). (These were made from one hundred separate runs of the network on the original data set of 50 members of Congress.) Even though average activation overlap between the two vectors is low because of a large number of shared inactive nodes, the amount of forgetting remains high. Notice, however, that all of the overlap occurs over a very small number of nodes. Thus, when the weights associated with these nodes change, this will disturb the prior representations for *both* Democrats and Republicans. Since only a few nodes are actually contributing activation, when the weights associated with one of these active nodes change significantly, very few other active nodes can "come to the rescue" and compensate for the changes to the other weights.

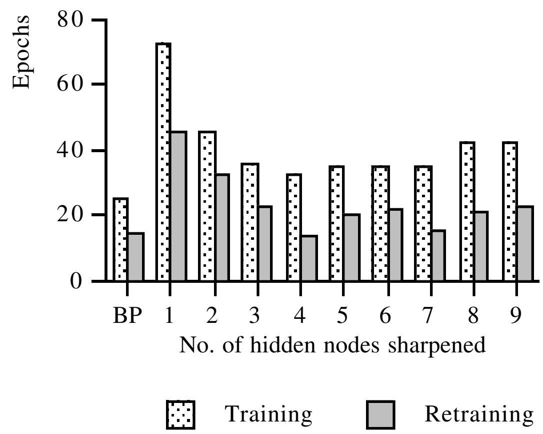

Figure 3. Training and retraining times for standard backpropagation versus sharpening

What we should see is that when activation patterns of representations are well distributed and well separated, relearning time should drop. As we can see in Figure 3, four-node sharpening, in which the representations are relatively well distributed (and reasonably well separated), causes retraining time to drop to what it was for standard backpropagation. Although the discussion is beyond the scope of this paper, it turns out that the amount of *separation* among competing representations is a very good predictor of catastrophic interference: the greater the separation, the less the forgetting. The reason that using the amount of activation overlap worked in some cases to predict catastrophic forgetting is quite simple: high representational separation implies low overlap (but not necessarily vice-versa). So in many cases, reducing overlap at the hidden layer appeared to be the cause of reduced catastrophic interference, but in reality, representational separation *also* increased, the latter being the actual cause for improved performance on catastrophic forgetting. As the Voting Records database example shows, decreasing overlap by using sparse representations is not always sufficient to reduce catastrophic forgetting; increased *separation* of representations is. (In fact, any time the non-zero activation level of some node is the same for all representations, that node becomes, in effect, a bias node, i.e., a node whose activation level, usually 1, remains unchanged for all inputs.)

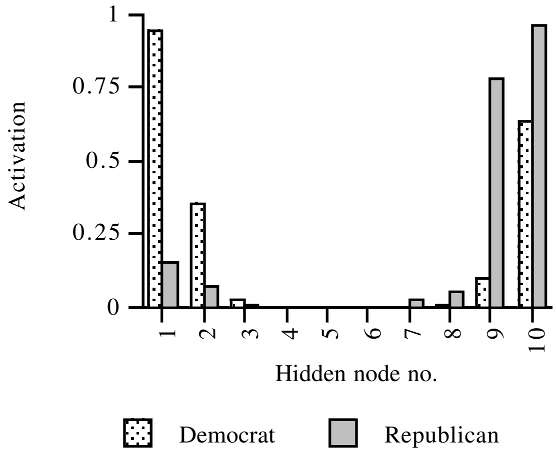

Figure 4: Two-node sharpened representations

Context biasing

Context biasing produces precisely the kind of representations that are both well distributed and well separated. The technique requires the network to be able to remember both the previous teacher pattern and the previous hidden-layer representation. When a new pattern is presented to the network, it is fed forward through the network, whereafter the weights are changed according to the standard backpropagation algorithm. The difference, measured in terms of the average Hamming distance, between the new teacher and the previous one is computed. (Given two k-bit vectors, the average Hamming distance between them is the number of bits that must be changed to transform one vector into the other divided by k.) The hidden-layer representation for the new association is then compared to the corresponding representation for the prior association. The new representation is then "separated" from the prior representation according to the following separation rule:

Modify the activation, A, of each node of the new representation as follows:

if $A_{new} \geq A_{previous}$
then $A_{new} = A_{new} + \alpha\beta(1 - A_{new})$;

if $A_{new} < A_{previous}$
then $A_{new} = (1 - \alpha\beta)\, A_{new}$

where
α = Hamming distance between the previous and the new teacher patterns;
β = biasing coefficient (usually 0.5 or 0.2)

This new "context-biased" representation (i.e., the "target" hidden-layer representation) is then "locked into" the input-to-hidden weights by backpropagating from the hidden layer to the input layer an "error signal" consisting of the difference between the "natural" hidden-layer representation and the corresponding "context-biased" (target) representation. This "locking-in" technique is discussed in (French 1991).

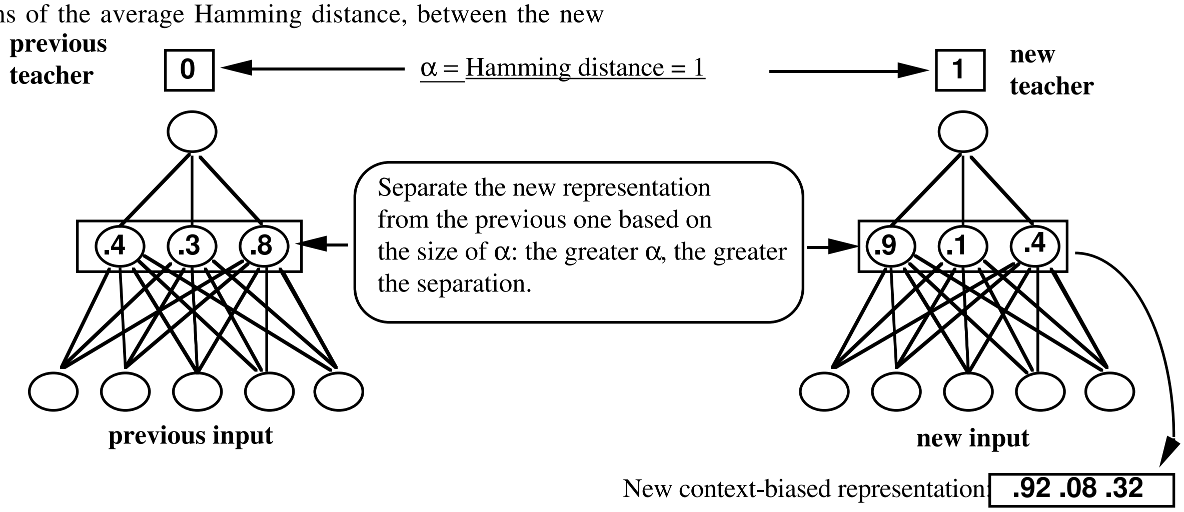

Figure 5. Modifying hidden-layer representations with context-biasing

Results

Figure 6 below shows hidden-layer representation profiles for Democrats and Republicans that are produced with standard backpropagation. (Data was collected over 100 runs of the program using a 16-10-1 network with a learning rate of 0.2 and momentum of 0.9.). Notice that, while the representations for Democrats and Republicans are well distributed, they are not particularly well separated.

On the other hand, when context biasing ($\beta = 0.2$) is used (Fig. 7), significantly different representation profiles result. Notice that the distributions for both Democrats and Republicans are not only well distributed across the entirehidden layer, but they are also well separated. If the above analysis is correct, we should therefore see a significant reduction of catastrophic forgetting with respect to backpropagation. As the results in Figure 8 show, that isindeed what happens. We get over a 50% reduction in relearning time when the context-biasing is used.

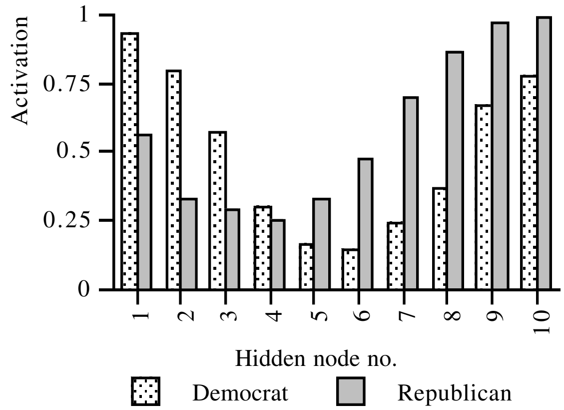

Figure 6. Hidden-layer representation profiles with backpropagation

For similar reasons — reduced interference during learning — context-biased networks also train up somewhat more quickly. Finally, preliminary work suggests that the reduction in relearning times is largely unaffected by the order of presentation of the data. In one experiment, all Republicans were presented to the network, followed by all Democrats and relearning times remained similar to those indicated in Figure 8.

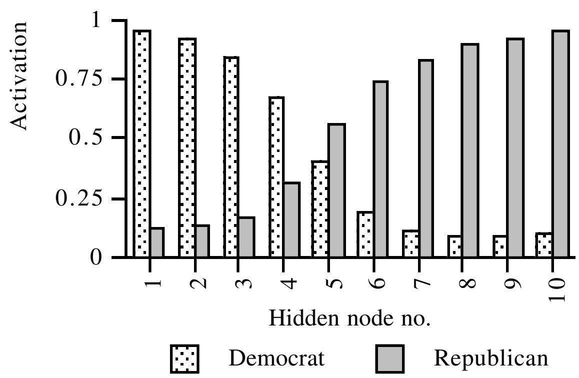

Figure 7 . Hidden-layer representation profiles with context-biasing

Conclusions

This paper addresses the fundamental problem of stability and sensitivity in learning. How can neural networks be designed to easily acquire new information without disturbing old, well-learned information — something we humans do so easily? Perhaps it will ultimately be necessary to develop tandem systems of neural networks with one "neocortical" network for storing well-rehearsed concepts and another "hippocampal" network to handle new input and to serve as a teacher to the neocortical network. French (1994) has developed such a two-system network and McClelland, McNaughton, & O'Reilly (1994) have also argued for the necessity of this type of two-tiered system.

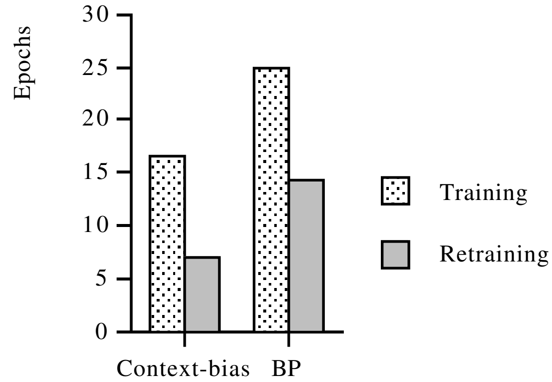

Figure 8. A significant (> 50%) reduction in forgetting with context biasing

However,the work presented in this paper suggests that at least some of the advantages of a two-network system can be achieved in a single network by using context biasing to appropriately constrain the hidden-layer representations during learning. In particular, context biasing produces representations that are both well distributed and as orthogonal as possible, thereby significantly reducing catastrophic forgetting.

Acknowledgments

I would like to thank Jim Friedrich and David Lutz for their contribution to this research, as well as David Chalmers and two reviewers for their helpful suggestions for the final draft of the paper.

References

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

French, R. M. (1994). "Catastrophic forgetting in connectionist networks: Can it be predicted, can it be prevented?" Summary of the NIPS-93 Workshop on Catastrophic Forgetting, In Cowan, J.D., Tesauro, G., & Alspector, J. (eds.) *Advances in Neural Information Processing Systems 6*. San Francisco, CA: Morgan Kauffmann (to appear).

Hetherington, P. & Seidenberg, M. (1989) "Is there 'catastrophic interference' in connectionist networks?" *Proceedings of the 11th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, 26-33.

Lewandowsky, S. & Goebel, R. (1991) "Gradual unlearning, catastrophic interference, and generalization in distributed memory models" University of Oklahoma Psychology Department Technical Report, presented at the *1991 Mathematical Psychology Conference*, Indiana University, Bloomington, IN.

Lewandowsky, S. & Shu-Chen Li (1993) "Catastrophic Interference in Neural Networks: Causes, Solutions, and

Data" in *New Perspectives on interference and inhibition in cognition* F.N. Dempster & C. Brainerd (eds.). New York, NY: Academic Press (to appear).

McClelland, J., McNaughton, B., & O'Reilly, R., Why there are complementary learning systems in the hippocampus and neocortex. CMU Tech Report PDP.CNS.94.1, March 1994.

McCloskey, M. & Cohen, N. J. (1989). "Catastrophic interference in connectionist networks: The sequential learning problem" *The Psychology of Learning and Motivation*, **24**, 109-165.

McRae, K. & Hetherington, P. (1993) "Catastrophic interference is eliminated in pretrained networks" *Proceedings of the 15h Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Murphy, P. & Aha, D. (1992). UCI repository of machine learning databases. Maintained at the Dept. of Information and Computer Science, U.C. Irvine, Irvine, CA.

Murre, J. (1992) "The effects of pattern presentation on interference in backpropagation networks" in *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

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