Reassessment of catastrophic interference

Makoto Yamaguchi[CA]

Department of Educational Psychology, Waseda University, Tokyo 169-8050, Japan

[CA]Corresponding Author: yamag-psy@kurenai.waseda.jp

Received 6 July 2004; accepted 3 September 2004

Connectionist models using the back propagation learning rule are known to have a serious problem in that they exhibit catastrophic interference (or forgetting) with sequential training. After the model learns the first set of patterns, if the model is trained on another set of patterns, its performance on the first set dramatically deteriorates very rapidly. The present study reconsiders this issue with three sets of simulations. With orthogonal input vectors, interference can be reasonably mild. The number of hidden units was critical for the degree of interference, in contrast to suggestions of previous studies. Output coding scheme was found to be critical. The length of input lists also influenced the degree of interference. This study suggests that the interference problem has been overstated in the literature. *NeuroReport* 15:2423–2426 © 2004 Lippincott Williams & Wilkins.

Key words: Back-propagation; Catastrophic interference; Computational; Connectionism

INTRODUCTION

Connectionist models with a back propagation learning algorithm are known to have a serious problem with sequential learning. This problem was brought to the attention of researchers by two very influential seminal studies. McCloskey and Cohen trained the model sequentially [1]. The model first learned the first set of patterns well, and then in the next stage learned the second set without rehearsing the first set. Their results shocked researchers as they revealed that as soon as the second stage began the learning of the first stage was destroyed almost completely. Hence this phenomenon is referred to as catastrophic interference or forgetting. Ratcliff demonstrated similarly severe interference [2]. From these studies, a picture emerged in which interference is extremely rapid and complete in fairly broad conditions, and increasing the number of hidden units has no effects. Unfortunately, most researchers who mention the catastrophic interference problem only cite the two seminal articles and took them at face value, except specialists on this issue (for review see [3]).

Kortge claimed that interference was not severe with orthogonal input vectors [4], but Ratcliff also used orthogonal vectors in the simulations and concluded that the degree of interference was problematic [2]. This conflict must be resolved. This study reinvestigates the issue with most basic models, that is, pure gradient descent which does not include momentum or any modified learning algorithms.

There are different criteria by which to assess performance of the model [5]. Here, cosine and correlation are computed between the output vector and the teaching signal vector. In the simulations below, after stage 1 training, both measures easily exceeded 0.99. All the following simulations were replicated 20 times. Tests in simulations 1 and 2 included 4 patterns, thus cosine and correlation were averaged across 80 values.

Such continuous measures are not necessarily easy to interpret, especially for intermediate values. Therefore, some discrete measure will be helpful. In the present study except simulation 3, the teaching signal to the output layer is local coding, with one unit of activation 0.9 (or 1) and all the others of activation 0.1 (or 0). The performance is considered to be correct if the most active unit in the output layer matches the one with teaching signal of 0.9 (or 1). That is, in response to (0.1, 0.9, 0.1, 0.1), the output is judged correct if it is (0.3, 0.7, 0.4, 0.2). Recently, Yamaguchi [6] used this discrete measure and found that the degree of interference is not severe, with the model seldom making errors.

SIMULATION 1a, 1b

Networks with four architectures are used, which are 8-4-8, 8-8-8, 8-16-8, and 8-32-8 networks, where the three numbers representing numbers of input, hidden, and output units, respectively. They learned essentially autoassociation (Table 1). Note that this task is very similar to that of Ratcliff [2], who found significant interference in autoassociation with 4-2-4 network and (0,1) teaching signal.

The results for continuous measures are shown in Fig. 1. For the cosine measure, chance level (numerically estimated from the output of the model not having been trained at all) is not zero but around 0.6. Chance level is zero for correlation, which holds in subsequent simulations as well. The degree of interference was influenced by the number of hidden units, in contrast to suggestions of previous studies. Also intriguing is that dissociation was found between cosine and correlation. From cosine, the performance of the 8-4-8 network was even below chance. However, correlation suggests it retains stage 1 information.

In the light of discrete measure, the number of total correct answers was 4, 58, 80, and 80 for 8-4-8, 8-8-8, 8-16-8, and 8-32-8 networks, respectively. That is, the network produced the correct answers to all the input patterns in the latter two conditions. With the number of hidden units increased, interference was reduced on the basis of all the measures. Errors were inspected to see their nature. Inspection revealed that in all the errors, the model erroneously selected an output unit given the teaching signal of 0.9 in stage 2. Also the same held for all the following simulations, so the model responded with one of the stage 2 facts whenever an error occurred.

Simulation 1b was a replication of simulation 1a except that teaching signal was converted from (0.1,0.9) to (0,1) like that used by Ratcliff [2]. The use of (0,1) teaching signal is not without problems. For instance, weights increase unboundedly.

The results for continuous measures are shown in Fig. 2. For the cosine measure, chance level is around 0.35. Overall, performance deteriorated compared with simulation 1a. Again, dissociation of cosine and correlation was found. The performance of 8-4-8 network was substantially below chance with cosine but correlation was almost zero. The number of correct answers was 1, 29, 60, and 74 for 8-4-8, 8-8-8, 8-16-8, and 8-32-8 networks, respectively, which is again worse than the previous simulation. But the basic pattern of results was the same; interference was meliorated with increased number of hidden units.

Table I. Task for simulation Ia.

Stage	Input	Output
1	1 0 0 0 0 0 0 0	0.9 0.1 0.1 0.1 0.1 0.1 0.1 0.1
	0 1 0 0 0 0 0 0	0.1 0.9 0.1 0.1 0.1 0.1 0.1 0.1
	0 0 1 0 0 0 0 0	0.1 0.1 0.9 0.1 0.1 0.1 0.1 0.1
	0 0 0 1 0 0 0 0	0.1 0.1 0.1 0.9 0.1 0.1 0.1 0.1
2	0 0 0 0 1 0 0 0	0.1 0.1 0.1 0.1 0.9 0.1 0.1 0.1
	0 0 0 0 0 1 0 0	0.1 0.1 0.1 0.1 0.1 0.9 0.1 0.1
	0 0 0 0 0 0 1 0	0.1 0.1 0.1 0.1 0.1 0.1 0.9 0.1
	0 0 0 0 0 0 0 1	0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.9

The results were comparable with the number of learning trials increased several fold (data not shown). Also, the results were comparable with the use of cross-entropy as an error function instead of summed squares (data not shown). As an important aside, McCloskey and Cohen claimed their model was sometimes stuck in a local minimum [1], but that was probably not a local minimum but a plateau and could be easily escaped with the use of cross-entropy, which enables more efficient learning for (0,1) teaching signal.

These results suggest that the two influential seminal studies [1,2] were too pessimistic. Kortge [4] and Lewandowsky [7] (see also [8]) claimed that interference is not severe with orthogonal inputs, although the latter study

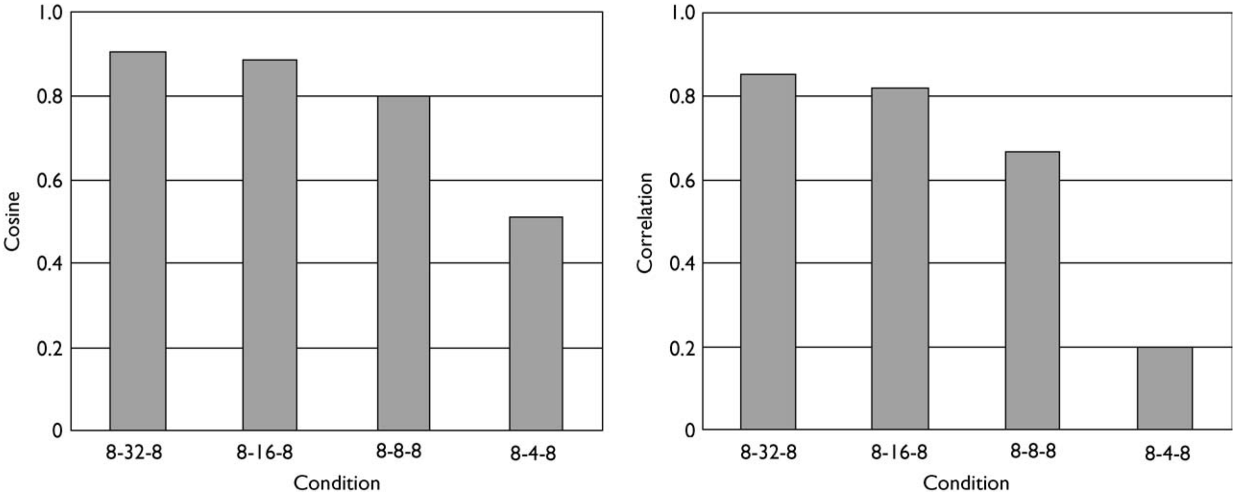

Fig. 1. Results of simulation Ia. Retention of stage I information in light of cosine (left panel) and correlation (right panel) is shown. Networks vary with respect to the number of hidden units.

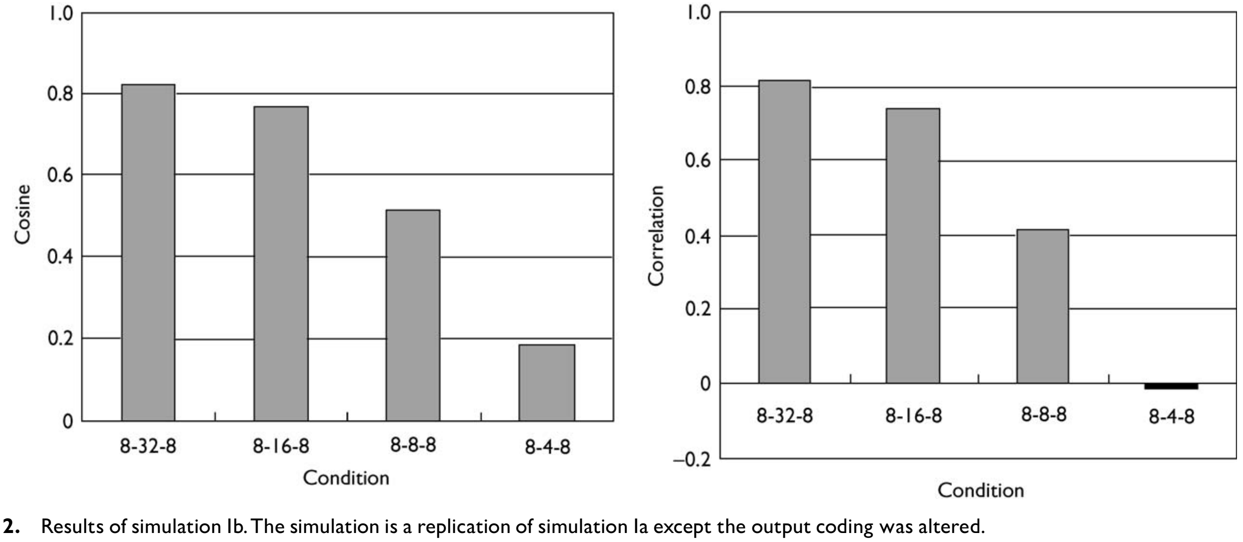

Fig. 2. Results of simulation Ib. The simulation is a replication of simulation Ia except the output coding was altered.

found gradual but complete forgetting. However, those studies adopted only a quantitative measure as a performance criterion. Therefore, by demonstrating errorless performance by the discrete criterion with enough hidden units, the present results extend the previous studies.

Some researchers (including [1]) have reported that catastrophically severe interference was not mitigated by increasing the number of hidden units. The present simulation however suggested that performance can be improved with sufficient number of hidden units. Many set the number of hidden units less than that of input units [2]. Results obtained with such networks are interesting from the perspective of statistical sciences, but this restriction leads to greater interference.

SIMULATION 2a, 2b

The input representation in the above simulations was localist. On each pattern presentation, only one input unit received input 1, and connections from the other units to hidden units were not altered because activation value 0 was multiplied. One can also obtain orthogonal vectors with distributed representation. For generality, further simulations were conducted with distributed and orthogonal input vectors.

Several previous researchers used input units with bipolar activation (1, −1) instead of (0,1) [6]. They randomly assigned 1 and −1 to units, which result in the expected inner product of zero (i.e., orthogonal). In this report, exactly orthogonal vectors were produced (Table 2). Here, for clarity, 1 is represented by ○ and −1 by ●. Technically, this is called an Hadamard matrix, whose rows (or columns) are exactly orthogonal to each other. Output representation was not altered. 8-4-8 network was investigated. Teaching signal of (0.1,0.9) was used (simulation 2a) as well as (0,1) (simulation 2b).

Results for continuous measures are shown in Fig. 3. For the discrete measure, the number of correct answers was 2 and 0 for simulations 2a and 2b, respectively. The results are almost identical to those of simulations 1a and 1b. Both the superior performance with (0.1,0.9) teaching signal and the dissociation of cosine and correlation were replicated. Compared to the previous simulations, very different input coding resulted in comparable degree of interference.

SIMULATION 3

In some respects these results are inconsistent with those of Yamaguchi [6]. Yamaguchi reported moderate interference for the network in which the number of hidden units is half the number of input units, but the 8-4-8 condition in the above results showed severe interference. This is an apparent contradiction that must be resolved.

The amount of information the network must retain may influence the degree of interference. Although this is plausible, previous studies did not investigate it explicitly. To explore this issue, further simulations were conducted. Now the network has 16 input and output units with 8 hidden units (16-8-16). Exactly orthogonal, distributed representation was again used (8-dimensional Hadamard matrix in Table 2 was expanded to 16 dimensions). There were three conditions: 4-4 condition learned four patterns first, followed by four patterns in the second stage; 8-8 condition learned eight patterns followed by another eight; and 8-4 condition learned eight patterns followed by four.

As these conditions differ in the number of parings that must be learned, if output coding was again localist, then two of the conditions would not utilize some output units, complicating interpretation. Therefore, output coding was altered to be distributed representation, derived from input vectors. However, to avoid changing the unit activation function that would be necessary to accommodate bipolar (1, −1) coding, 1 was translated to 0.9, and −1 was translated to 0.1. Although output vectors were straightforwardly derived from input vectors, they are not orthogonal to each other. In total, the test

Table 2. Input patterns used in simulations 2a, 2b.

○	○	○	○	○	○	○	○
○	○	○	○	●	●	●	●
○	○	●	●	○	○	●	●
○	○	●	●	●	●	○	○
○	●	○	●	○	●	○	●
○	●	○	●	●	○	●	○
○	●	●	○	○	●	●	○
○	●	●	○	●	○	○	●

○ represents 1 and ● represents −1.

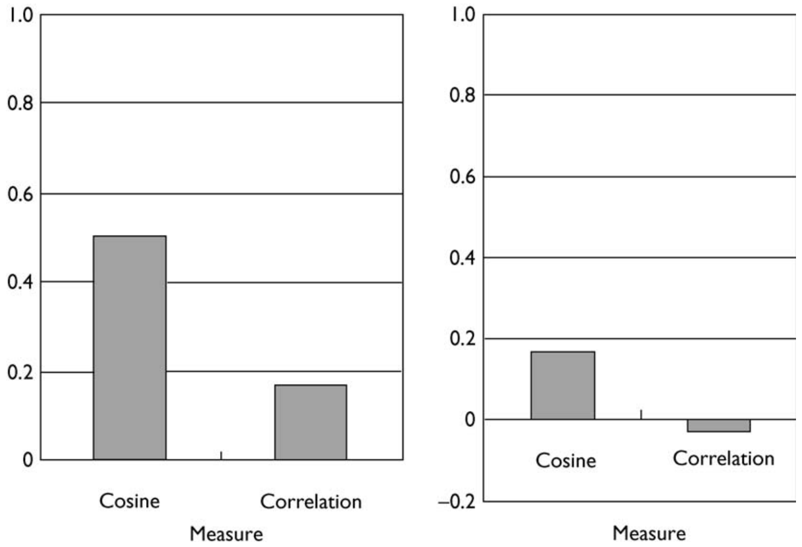

Fig. 3. Results of simulation 2a (left panel) and 2b (right panel). Distributed but orthogonal input coding was used for the 8-4-8 network.

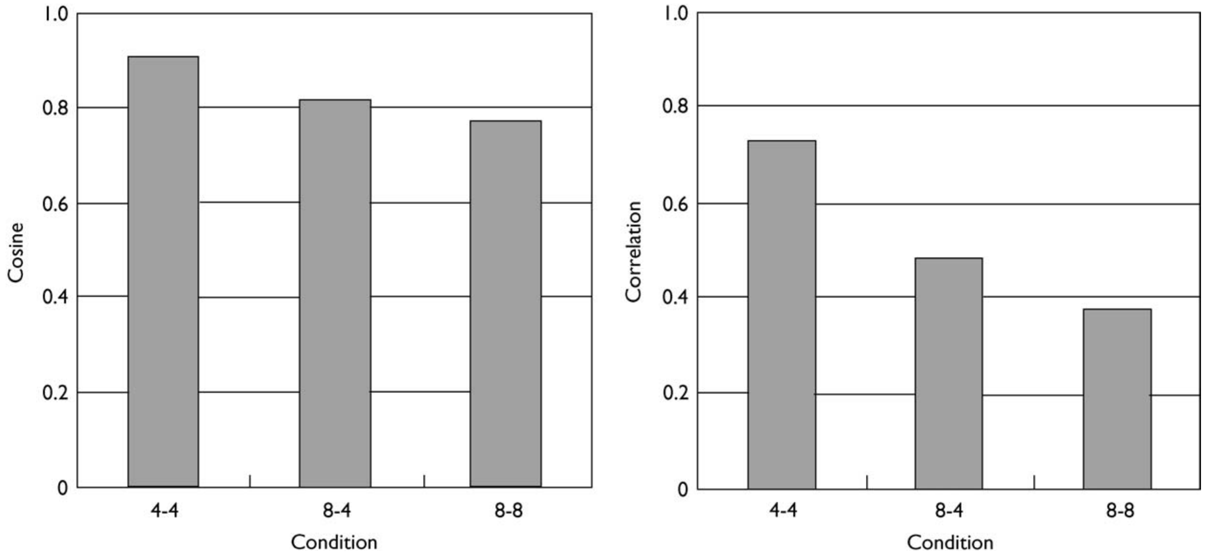

Fig. 4. Results of simulation 3. The 16-8-16 network learned various numbers of associations.

includes 80 probes in 4-4 condition and 160 probes in 8-4 and 8-8 conditions, across which cosine and correlation were averaged. Discrete measure was not used this time.

The results are shown in Fig. 4. For the cosine measure, the chance level is around 0.8. The results indicate that the number of patterns the model is given influences the degree of interference. Mild interference in [6] was conceivably in part due to this factor. Again dissociation was found between cosine and correlation. From cosine measure, 8-4 and 8-8 conditions are judged to retain no information for the first stage, but correlation suggests all three conditions retain the information.

CONCLUSION

When discussing the interference problem in connectionist models, most authors cite only the two seminal papers and claim that connectionist models always exhibit catastrophic interference in sequential learning. This article showed this understanding is not accurate, and suggests that many researchers rather overestimated the problem. Interference exhibited by models can be gradual and mild for orthogonal vectors. This was accomplished without any modifications of the learning rule. Although Kortge [4] suspected that the usually used asymmetrical unit activation function (which ranges from 0 to 1) results in greater interference, mild interference was obtained without changing the activation function. The author's claim is not that the interference problem was already resolved, but that it is wrong to always presume the catastrophic interference in connectionist models.

In contrast to the claims of some of earlier reports, the number of hidden units is critical and the interference is mitigated with increased number. Whether input is asymmetrical (0,1) or bipolar (1, -1), the degree of interference is comparable as far as it is orthogonal. With respect to output coding, however, (0,1) teaching signal results in greater interference than (0.1,0.9) signal does. The amount of information the network must retain is also shown to be critical, the longer list leading to the greater interference.

This research revealed dissociation of two measures, cosine and correlation. Some results were interpreted differently as below chance, just at the chance level, or above chance, by cosine and correlation. At this stage, its reason and which is more appropriate measure are unclear, but at least correlation may be more sensitive measure because chance level is zero. Thus, this article leaves one problem for future researchers

REFERENCES

1. McCloskey M and Cohen NJ. Catastrophic interference in connectionist networks: the sequential learning problem. In: Bower GH (ed.) *The Psychology of Learning and Motivation*. Vol. 24. New York: Academic Press; 1989, pp. 109–165.
2. Ratcliff R. Connectionists models of memory: constraints imposed by learning and forgetting functions. *Psychol Rev* 1990; **97**:285–308.
3. French RM. Catastrophic forgetting in connectionist networks. *Trends Cogn Sci* 1999; **3**:128–135.
4. Kortge CA. Episodic memory in connectionist networks. *Proceedings of the 12th Annual Conference of the Cognitive Science Society* 1990; 764–771.
5. Goebel RP and Lewandowsky S. Retrieval measures in distributed memory models. In: Hockley WE and Lewandowsky S (eds). *Relating Theory and Data: Essays on Human Memory in Honor of Bennet B. Murdock*. Hillsdale, NJ: Lawrence Erlbaum; 1991, pp. 509–528.
6. Yamaguchi M. Are multi-layer back propagation networks catastrophically amnesic? *Scand J Psychol* 2004; **45**, in press.
7. Lewandowsky S. Gradual unlearning and catastrophic interference: a comparison of distributed architectures. In: Hockley WE and Lewandowsky S (eds). *Relating theory and data: essays on human memory in honor of Bennet B. Murdock*. Hillsdale, NJ: Lawrence Erlbaum; 1991, pp. 445–476.
8. Lewandowsky S and Li S-C. Catastrophic interference in neural networks: causes, solutions, and data. In: Dempster FN and Brainerd C (eds). *Interference and Inhibition in Cognition*. San Diego: Academic Press; 1995, pp. 329–361.

Acknowledgements: This research was funded by Grant-in-Aid for JSPS Fellows 14-03153. The simulations were conducted using software tlearn, freely available at (http://crl.ucsd.edu/innate/tlearn.html). The authors are grateful for the developers of the software.