Understanding Failures of Learning: Hebbian Learning, Competition for Representational Space, and some Preliminary Experimental Data

James L. McClelland[1,2], Adam Thomas[1,3], Bruce D. McCandliss[1,4] and Julie A. Fiez[1,4]

Center for the Neural Basis of Cognition[1],

Department of Psychology, Carnegie Mellon University[2],

Department of Neuroscience, University of Pittsburgh[3],

Learning Research and Development Center and

Department of Psychology, University of Pittsburgh[4]

Send Correspondence to:

James L. McClelland
Center for the Neural Basis of Cognition
115 Mellon Institute
4400 Fifth Avenue
Pittsburgh, PA 15213
jlm@cnbc.cmu.edu
(412)-268-3157 (Voice) / (412)-268-5060 (Fax)

Introduction

The availability of powerful learning algorithms such as back-propagation has created a situation in which we now know how to teach neural networks many complex things. Models that use back propagation have been used to account for development and learning in a wide range of task situations. For example, there are successful models of the acquisition of word reading skill (Sejnowski & Rosenberg, 1987; Seidenberg & McClelland, 1989; Plaut, McClelland, Seidenberg, & Patterson, 1996), of physical knowledge such as object permanence (Munakata, McClelland, Johnson, & Siegler, 1997), and of conceptual knowledge (Hinton, 1989; Rumelhart, 1990) such as kinship relations and the natural kind hierarchy. These models raise the question, why is it that we sometimes fail to learn from experience?

Two cases of failure of learning motivated the present analysis:

1. Many models, including the model of McClelland, McNaughton, and O'Reilly (1995), assume that after the onset of amnesia, gradual learning in spared neocortical areas is possible. This assumption is based on the fact that amnesic subjects show considerable spared learning ability, particularly when given repeated exposure to items, and plays a prominent role in most models ■of temporally graded retrograde amnesia McClelland, McNaughton, and O'Reilly (1993; Alvarez & Squire, 1994; Milner, 1989). Yet they are virtually completely incapable of learning a set of arbitrary paired associates using standard paired-associate learning methods, even with massive repetition. This is mentioned in passing in several case reports, but detailed documentation is lacking. However, a similar failure is reported by Gabrieli, Cohen, and Corkin (1988). They tried to teach HM the meanings of eight new words, one of which he already knew in advance. Through hundreds of trials over many sessions with several task variants, he failed to learn the meanings of any of the seven words he did not already know.

2. Even in normal adults, there can be cases of failure of learning. For example, when Japanese Adults come to the United States, they often have great difficulty discriminating between /r/ and

/l/; while there is some evidence of slight improvement over time, it is very gradual, and difficulties can persist indefinitely. Yet adults are capable of learning many new skills, and indeed it seems likely that the cortical mechanisms thought to underlie spared learning in amnesics are available for skill learning in normal subjects. Why then are perceptual discriminations between sounds not contrasted in one's own native language so difficult to acquire?

One idea that could account for these difficulties comes from a consideration of the Hebbian learning rule. According to Hebb, if one neuron takes part in firing another, the strength of the connection between them will be increased. What this means is that if an input elicits a pattern of neural activity, Hebbian learning will tend to strengthen the tendency to elicit the same pattern of activity on subsequent occasions. That is, if learning in the brain is Hebbian, then learning will tend to strengthen whatever response the brain makes to its inputs. If the response is useful and constructive, the brain will learn to reinforce it. If the response is inappropriate or undesirable, Hebbian learning will still tend to reinforce it. This leads to the suggestion that many failures of learning in adulthood may reflect a paradoxical tendency of the mechanisms of learning to reinforce inappropriate or undesirable responses.

We can now examine why paired associate learning may be difficult in amnesics. The subject receives a list of, say, 12 word pairs (including, for example, LOCOMOTIVE-DISHTOWEL and TABLE-BANANA, among others). After a slight delay, the experimenter presents the first word in one of the pairs, and asks the subject to recall the word that was previously paired with it in the experiment. Due to the subject's amnesia, he may not remember even that there was a list of word pairs. Nevertheless, as is standard in paired-associate learning, the subject is encouraged to guess a response. Given the arbitrary pairing of the words, TABLE is unlikely to come to mind in the context of BANANA as a cue, and so the stimulus is likely to elicit some other response. If learning is Hebbian, is this response that will be strengthened, thereby leading to interference.

There is experimental support for the idea that forcing amnesics to make their own responses to items leads to interference. (Baddeley & Wilson, 1994) contrasted two conditions for teaching subjects to recall a particular whole word (e.g. QUOTE) from a part-word cue (e.g. QU). In the errorless condition, the experimenter said: ”I’m thinking of a word beginning with QU. The word is QUOTE. Please write it down.” In the errorful condition, the experimenter said. ”I am thinking of a word beginning with QU. Can you guess what it is?”. Subjects generally made several incorrect guesses, and in fact the experimenter could switch the ”correct” answer to ensure that no correct guesses were made on the first occasion. After the guessing the experimenter said ”The word I was thinking of is QUOTE, please write it down.” Thus in both conditions, subjects wrote down the experimenter’s word at the end of the presentation. This procedure was repeated three times with several different words in each condition. Subsequently, subjects were tested for their ability to remember the experimenter’s words. The amnesic subjects scored about 3070them to guess their own responses produced massive interference. Control subjects did far better in both conditions and showed much less interference from their own guesses. Another experiment making similar points was reported by Hayman, Macdonald, and Tulving (1993).

Modeling Studies

The focus of the research we report here has been the failure of Japanese adults to learn the discrimination between /r/ and /l/. We have taken a two-pronged attack, combining computational modeling with experimental studies on Japanese adults who show considerable difficulty discriminating /r/ and /l/.

Insert Figure 1 about here.

The modeling work (Thomas & McClelland, 1997) began with an effort to demonstrate how a Hebb-like learning mechanism could lead to failure to learn to discriminate two similar inputs (abstract proxies for /r/ from /l/), once that ability had been lost by ’rearing’ the network in an environment providing only a single input in that region of perceptual space. For this purpose we used a variant of the Kohonen network architecture (Figure 1). There were two layers, each with 49 units, arranged in a 7x7 array. These layers were called the ’input’ and the ’representation’ layer, respectively. Initial random feed-forward projects were loosely topographic. On presentation of an input, the representation unit receiving the strongest net input was chosen as the winning representation unit, and it and its neighbors were assigned activation values equal to a Gaussian function of distance from the winner. Weights coming into representation units were then adjusted according to a variant of the competitive learning rule:

$$\Delta w_{rs} = \varepsilon a_r (a_s - w_{rs}) \qquad (1)$$

Here a_r is the activation of the receiving or representation unit, a_s is the activation of the sending or input unit, and w_{rs} is the weight to the former from the latter. This rule has a Hebbian component (the product $a_r a_s$) together with a tendency for weights to decay in proportion to the product of the activation of the receiving unit and the current value of the weight (the product $a_r w_{rs}$).

Insert Figure 2 about here.

Inputs consisted of Gaussian blobs of activity. Two training conditions were used. In both, there were four *corner* inputs (Figure 2A). These occupy the four corners of the input space, and correspond to background phonemes. In one training condition (the ”English-like” training condition) there were two additional *overlapping* inputs (Figure 2B), proxies for /r/ and /l/ respectively. Given the parameters used, more than 90successfully learned to assign distinct representations to the two overlapping stimuli, just as most children in English-speaking countries naturally learn to discriminate stimuli in their native language. In another training condition (the ”initially Japanese-like” training condition), there was initially just the four corner inputs and just one other, *centered* input (Figure 2C) between the two overlapping inputs used in the English-like conditions. After 300 epochs, the networks were switched from the Japanese-like to environment to the English-like environment. In this case, all networks learned initially to assign a single representation to the Japanese-like, centered input. Crucially, *none* of these networks subsequently learned to to assign different representations to the two overlapping English-like inputs. They retained their tendency to treat these inputs as the same, even though the mechanisms of plasticity operate without any changes throughout the simulation.

Thus far our work supports the idea that discriminations that can be learned if the distinction is present in a network’s initial environment may not be learned when the distinction is not introduced until after the network’s response tendencies become established. This provides one way of accounting for the the loss of plasticity seen in Japanese Adults. Obviously, other factors could be at work, including possibly a general reduction in plasticity with age.

If the mechanisms considered here are even part of the story, they predict that we may be able to induce plasticity in Japanese adults. First, we consider the use of exaggerated inputs as a method for inducing plasticity. To illustrate this in the simulation, we added two additional inputs to the English-like environment. These were exaggerated versions of the /r/ and /l/-like stimuli. Networks that failed to learn to discriminate the /r/ and /l/-like stimuli in the initially Japanese condition learned to discriminate these stimuli in only a few epochs after the exaggerated stimuli were included in the training set ((Figure 2D).

The idea that the use of exaggerated stimuli could induce plasticity is consistent with the findings reported by Merzenich, Tallal, and their colleagues Merzenich, Jenkins, Johnson, Schreiner, Miller, and Tallal (1996; Tallal, Miller, Bedi, Byma, Wang, Nagaraja, & others, 1996).

They showed that they could remediate children with language impairments when they used a training regime that exaggerated contrasts between plosive stops and other sounds differing by rapid transitions (See also Alexander & Frost, 1982). We wanted to show that plasticity was still present in adults, and to test the role of exaggeration. For this purpose, McCandliss, Fiez, Conway, Protopapas, and McClelland (1998) developed a set of two speech continua, one spanning from ”rock” to ”lock” and one spanning from ”road” to ”load”. Starting with natural speech tokens generated by a native English speaker, eighty-item continua were constructed for each contrast, ranging from highly exaggerated tokens of ”lock” or ”load” to highly exaggerated tokens of ”rock” or ”road”. University of Pittsburgh Undergraduates showed very clean categorization of the stimuli on each continuum; in each case, only 10 steps on the continuum lay in a gray zone, separating those items the native English speakers reliably heard as /l/ from an item they reliably heard as /r/.

Experimental Investigations

Eight subjects whose initial discrimination of /r/ and /l/ stimuli was quite poor were tested in the *adaptive* condition of this experiment. Each subject was trained on one of the two continua. Highly exaggerated tokens of /r/ and /l/ stimuli were used initially, near the extreme ends of the continuum. The two selected stimuli were presented in random order, and the subject was simply required to press one button if the stimulus began with /r/ and another if it began with /l/. Whenever the subject made an error, the task was made easier, by replacing the stimulus with a the next more exaggerated one, until the extremes of the continuum was reached. Whenever the subject performed correctly on eight trials in a row, the task was made easier, by replacing the /r/ or the /l/ with the next less exaggerated item. Half the subjects received feedback after each trial, and half received no feedback. All of the subjects showed substantial improvement within three twenty-minute sessions, and all showed marked improvement compared to pre-test performance in a subsequent post-test.

Eight other subjects selected according to the same criteria participated in the *set training* condition of the experiment. For this condition, the /r/ and /l/ stimuli just at the edge of the native English speaker’s gray zone were used throughout the experiment. Otherwise the experiment was identical to the adaptive condition, with half of the subjects receiving feedback and half receiving no feedback.

We had initially expected that the subjects would generally fail to learn the discrimination, but this expectation was only partially confirmed. The two out of the four subjects in the no feedback condition whose performance was initially the worst did fail to learn. If anything, these subjects became less able to distinguish the stimuli over the course of the experiment, in accordance with the Hebbian hypothesis. However, the other two subjects in the no-feedback condition, whose initial ability to distinguish the /r/-/l/ stimuli was somewhat better, showed rapid learning, with strong gains on the post-test. These findings suggest that as long as the stimuli are even only partially discriminable, the mechanisms of learning will successfully pull them apart.

Initially, we were puzzled by the fact that these subjects could learn so rapidly using very difficult stimuli without feedback. If they could learn this rapidly in our experiment, why had they not mastered the discrimination from exposure to natural speech? A possible explanation comes from an aspect of the model that we now feel may be at least as important as the use of Hebbian learning. This is the fact that, in a Kohonen network, patterns compete for space. The outcome of the competition depends on similarity, frequency of presentation, and existing conditions. Under natural conditions, we suggest that /r/ and /l/ must compete for space with many other stimuli. This happens in our simulations, where corner inputs compete for space with the overlapping inputs. Under these circumstances, if existing conditions are such that the overlapping stimuli are treated the same, the competition from the corner stimuli helps to maintain this. However, if training is focused only on the overlapping stimuli, the model will learn to separate them. Any initial difference in the response to the stimuli will be capitalized on very rapidly, leading to a rapid separation of the response to these two inputs. This is consistent with the conditions of our experiment, in which subjects were allowed to focus only on the distinction between /r/ and /l/ in a single contrast. Those subjects who had some initial ability to discriminate the stimuli rapidly learned to distinguish them.

Conclusions

The research reviewed above suggests that our initial suggestion that learning may rely on a Hebbian process may only be part of the story. It appears that competition for space in representations may also be relevant to understanding cases in which learning fails. To test this, we currently plan additional experiments in which we will vary the number of different stimuli the subjects must distinguish. According to the model, if several other stimuli ending in say ”_ock” are included at the same time in addition to ”rock” and ”lock”, the competition from the others for representational space should greatly retard learning of the /r/-/l/ discrimination.

Thus far we have considered only the results for the subjects who received no feedback in the set training condition of the experiment. Four other subjects in this condition did receive feedback, and all four of them showed rapid improvement. This indicates that in fact, outcomes can make a difference to learning, and that the Hebbian account of learning is at best incomplete. Current work in our group is exploring ways in which the existing model might be elaborated to take account of these aspects of the findings of this experiment.

While feedback does appear to play a role, it is our view that such feedback is very rarely available in natural learning situations. True, a Japanese adult may be able to use context to determine whether an English speaker intends to refer to a rock or a lock. However, such context cannot unambiguously tell us that the words for rocks and locks are different. There are clearly words within both English and Japanese that sound the same but refer to completely different things in different contexts (consider the English words ”bat” and ”ball”, which each of which have at least two apparently completely unrelated meanings). Thus our future efforts will consider how feedback might modulate and enhance learning that might otherwise depend crucially on the competitive and Hebbian characteristics found in the Kohonen net.

References

Alexander, D. W., & Frost, B. P. (1982). Decelerated synthesized speech as a means of shaping speed of auditory processing of children with delayed language. *Perceptual and Motor Skills*, *55*, 783–792.

Alvarez, P., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: A simple network model. *Proceedings of the National Academy of Sciences, USA*, *91*, 7041–7045.

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia*, *32*, 53–68.

Gabrieli, J. D. E., Cohen, N. J., & Corkin, S. (1988). The impaired learning of semantic knowledge following bilateral medial temporal-lobe resection. *Brain and Cognition*, *7*, 157–177.

Hayman, C. A. G., Macdonald, C. A., & Tulving, E. (1993). The role of repetition and associative interference in new semantic learning in amnesia: A case experiment. *Journal of Cognitive Neuroscience*, *5*, 375–389.

Hinton, G. E. (1989). Learning distributed representations of concepts. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 46–61). Oxford, England: Clarendon Press.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59–69.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*, 1464–1480.

McCandliss, B. D., Fiez, J. A., Conway, M., Protopapas, A., & McClelland, J. L. (1998). Eliciting adult plasticity: both adaptive and non-adaptive training improves japanese adults identification of english /r/ and /l/. *Society of Neuroscience Abstracts*, *24*, 1898.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1993). Why do we have a special learning system in the hippocampus?, (Abstract 580). *Bulletin of the Psychonomic Society*, *31*, 404.

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

Merzenich, M. M., Jenkins, W. M., Johnson, P., Schreiner, C., Miller, S. L., & Tallal, P. (1996). Temporal processing deficits of language-learning impaired children ameliorated by training. *Science*, *271*, 77–81.

Milner, P. (1989). A cell assembly theory of hippocampal amnesia. *Neuropsychologia*, *27*, 23–30.

Munakata, Y., McClelland, J. L., Johnson, M. H., & Siegler, R. (1997). Rethinking infant knowledge: Toward an adaptive process accout of successes and failures in object permanence tasks. *Psychological Review*, *104*, 686–713.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks* (pp. 405–420). San Diego, CA: Academic Press.

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

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce english text. *Complex Systems*, *1*, 145–168.

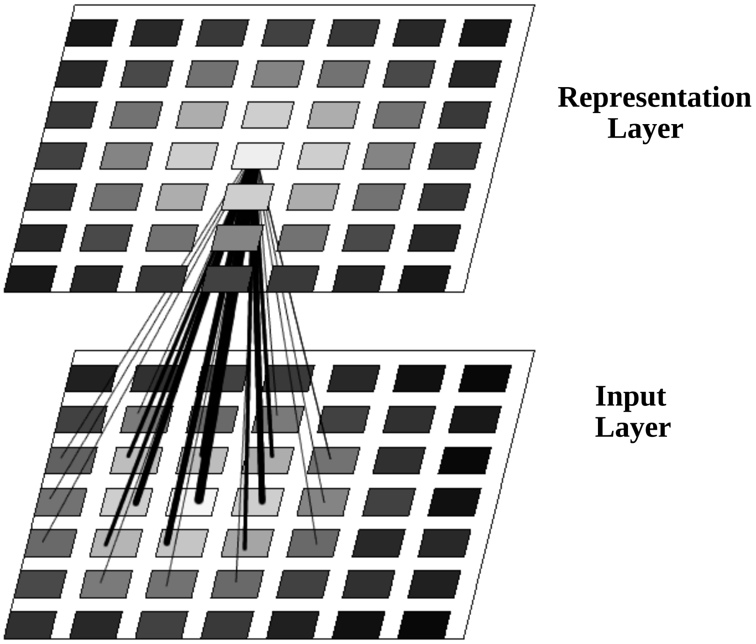
Tallal, P., Miller, S. L., Bedi, G., Byma, G., Wang, X., Nagaraja, S. S., Schreiner, C., Jenkins, W. M., & Merzenich, M. M. (1996). Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science*, *271*, 81–84.

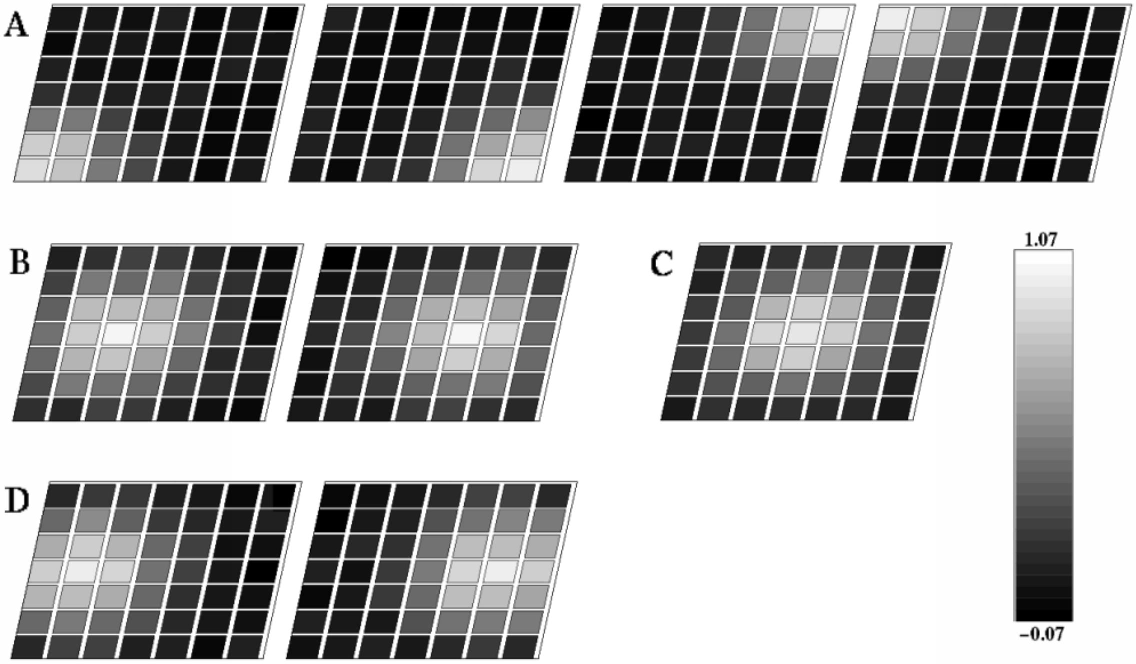
Thomas, A., & McClelland, J. L. (1997). How plasticity can prevent adaptation: Induction and remediation of perceptual consequences of early esxperience (abstract 97.2). *Society for Neuroscience Abstracts*, *23*, 234.

Figure Captions

Figure 1. A: The Self-Organizing Map network used in the simulations, adapted from (Kohonen, 1982, 1990)

Figure 2. Examples of the input patterns used in training and testing the network. (B) The four corner inputs. (C) The two overlapping inputs. (D) The single central input. (E) Exaggerated versions of the overlapping inputs used in ”remediation” of the network.


Representation Layer
Input Layer


A
B
C
D
1.07
−0.07