36

Connectionist Models of Memory

JAMES L. MCCLELLAND

The study of memory has been pursued within many different paradigms, and memory has been thought of in many different ways. Often memory has been viewed as a storehouse of items. Items are created, and then filed away as if they were books in a library. Storage, loss, organization, and retrieval of memories have all been considered from this point of view. Other theories (e.g., Anderson, 1983) hold that memory is a network of nodes with associative connections among them. Typically such theories propose some primitive or elemental nodes, and additional nodes that serve to group or organize collections of other nodes via associative links. This approach provides a basis for understanding how operations at the time of storage (creation of nodes and links) affect the success of later operations (e.g., retrieval of memories following links from node to node). In all these theories, memory consists of a set of items, together with an addressing scheme that allows them to be accessed.

The view we will consider in the present chapter begins with a completely different way of thinking about memory. This view addresses phenomena that have motivated the sorts of theories mentioned above, but in doing so it lets go of the idea that items are stored in memory as such. Instead the fundamental idea is that what is stored in memory is a set of changes in the instructions neurons send to each other, affecting what patterns of activity can be constructed from given inputs. When an event is experienced, on this view, it creates a pattern of activity over a set of processing units. This pattern of activity is considered to be the representation of the event. The formation of this pattern of activity provides the trigger for the creation of the instructions. The set of instructions is then stored in the connections among the units, where it is available for use in the construction of subsequent patterns of activity. Under some circumstances—for example, when the constructive process takes place in response to a recall cue—the cue may result in the construction of a pattern of activation that can be viewed as an attempted reconstruction of the pattern that represented the previously experienced event. Such a reconstructed representation corresponds to a recollection. The patterns themselves are not stored, and hence are not really "retrieved": recall amounts not to retrieval but to reconstruction.

The goal of the present chapter is to help the reader understand and appreciate the connectionist modeling framework, which embodies this view of memory. There are other modeling frameworks that take a very similar perspective, including the matrix model of Humphreys, Bain, and Pike (1989); the convo-

lutional model (TODAM) of Murdock (1982); and the composite holographic model of Metcalfe (1990). What sets the connectionist model apart from the others is the explicit use of the idea that the knowledge used in information processing—the instructions to which we referred just above—consists of the values, or *weights* on the connections among simple, neuronlike processing units. The approach thus taps into a vast neuroscience literature on the synaptic basis of learning and memory (see, e.g., McNaughton, 1993, for a review) as well as a complementary literature on the computational analysis of artificial neural networks (see the handbook by Arbib, 1995).

On the whole, the connectionist models have been motivated by robust and general patterns in data and by basic observations about the nature of human memory, and the focus has been on exploring general principles and understanding basic model characteristics. As the general principles become clearer, more detailed models of the quantitative aspects of experimental findings are likely soon to follow.

Local and Distributed Representation in Connectionist Networks

Some early connectionist models of memory (e.g., McClelland, 1981) made use of local representations, in which a single unit is used for each item. Typically the features of the item are represented by other units, to which the item unit is connected. Often these models can appear similar to symbolic associative models like Anderson's (1983), in that complex items are represented by units with connections to other, more elemental units. What differs is that in the connectionist model several items can contribute to the activation of elements, thus producing such phenomena as blend errors (Loftus, 1991; Nystrom & McClelland, 1992) and prototype effects (Posner & Keele, 1968).

A key question is whether the representation of knowledge in the brain is local in this way. The alternative—one that can be traced from Lashley (1950) through Willshaw (1981), Anderson (1973; Anderson, Silverstein, Ritz, & Jones, 1977) and Hinton (1981)—is an idea called *distributed representation*. This is the idea that the representation of an item is not associated with the activation of a single unit, but with the pattern of activation over a set of units; and the further idea that the knowledge underlying the ability to activate a particular pattern is not associated with the weights coming in and out of a single unit, but is superimposed in the same set of connection weights that encode the knowledge underlying other patterns.

A simple illustration of this idea at work (McClelland, Rumelhart, & Hinton, 1986) considers knowledge that allows one pattern to be associated with another. One pattern might represent the aroma of something, perhaps a rose, and the other its appearance. Figure 36.1A shows a network that allows the input pattern—the smell of the rose—to produce the corresponding output pattern. We imagine that this association is learned by experiencing the aroma and the visual appearance of the rose at the same time, producing the indicated patterns of activation on a group of units representing the aroma and on other units representing the appearance. Learning occurs by incrementing the weight to appearance unit i from aroma unit j in proportion to the product of the activations of the two units. This rule is often called a "Hebbian" learning rule, but to avoid confusion with a later discussion of Hebb's actual proposals, we will here describe this rule as a "coincident activation" learning rule. In the example, the weights all started at 0 and were incremented by the product of the unit activations times a rate constant ε (here set to 0.25):

$$\Delta w_{ij} = \varepsilon a_i a_j \qquad (1)$$

Now that we have stored the association, we can reconstruct the appearance from the aroma by presenting the pattern representing the aroma, and then setting the activations of the appearance units based on these activations and the learned weights. For simplicity in this case the activation is set to the net or summed input to the unit,

$$net_i = \Sigma_j a_j w_{ij}. \qquad (2)$$

The reader can check that the resulting activations exactly match the appearance pattern present at the time the increments to the connection weights were computed.

One can follow the same procedure to store a second association, perhaps between the aroma and the appearance of a steak, in the same set of weights. The same learning rule is used, and the increments for the steak are added to the weights for the rose. The second

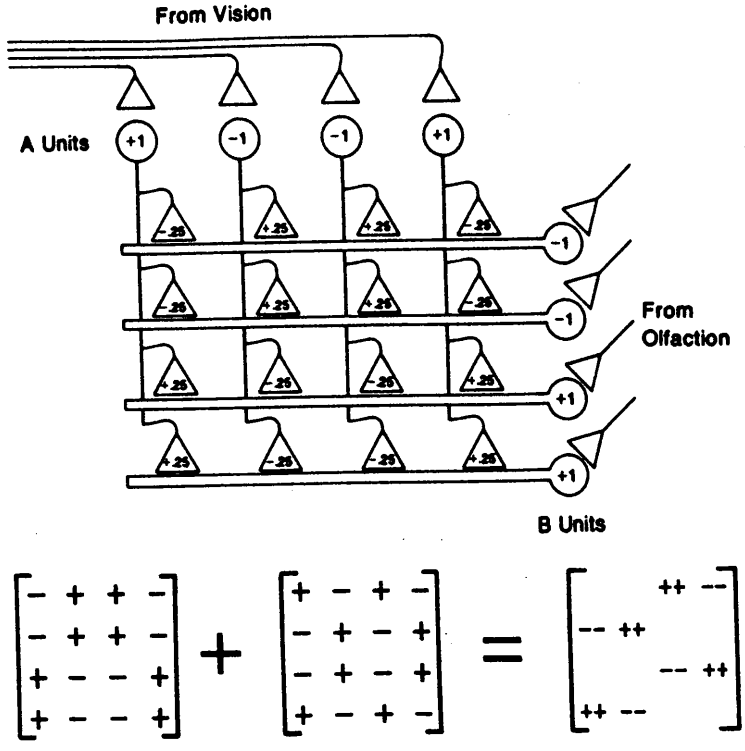

Figure 36.1 The simple associative network model used to illustrate how knowledge of two associations can be stored in the same set of weights. (A) The network with weights storing the single association between the given input and output patterns. (B) On the left, a schematic representation of the weight matrix appropriate for the same association shown in (A) and a second matrix appropriate for a second association between a different input pattern and a different output pattern. These two can be summed, creating the matrix shown on the right. Here + represents a value of +.25 and each – represents a value of –.25; ++ represents +.50, and – represents –.50. Adapted from McClelland et al. (1986), figures 12, 13, and 14, permission pending.

set of increments and the resulting summed values of the weights are shown in figures 36.1B and 36.1C. The reader can check that if one presents the rose aroma to the input, the output is the appearance of the rose, and if one presents the steak aroma, the output is the appearance of the steak.

A Distributed Auto-Associator Model of Memory

The above example demonstrates that the "memory trace" for something need not maintain a separate identity. Rather, the memory trace may be nothing more than a set of adjustments or increments to a large ensemble of widely distributed elements, the connection weights. The distributed memory model of McClelland and Rumelhart (1985) incorporates this idea. The model draws heavily on the work of James Anderson (1973; Anderson et al., 1977; Knapp & Anderson, 1984), and is a member of a class of models known as *auto-associator* or *attractor network* models (see figure 36.2). Instead of associating a pattern with another pattern, such models associate each pattern with itself. The network allows external inputs to all of the units and provides connections to each unit from every other

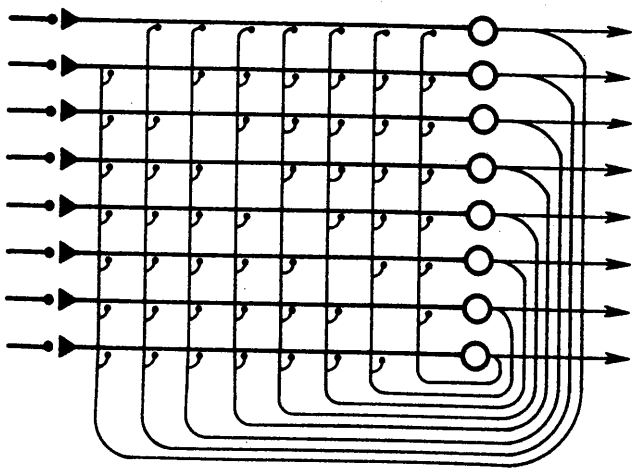

Figure 36.2 The auto-associator network from McClelland and Rumelhart (1985) Figure 1.

unit. The activation of a unit is determined by the net input, which consists of the external input plus the input coming to each unit from each of the other units via the weighted connections. The model uses an error correcting learning rule. The change in the weight to a particular unit i from another unit j is equal to:

$$\Delta w_{ij} = \varepsilon(ext_i - int_i)a_j \qquad (3)$$

This is often called the *delta* rule; learning is driven by the difference between one signal and another signal arriving at the receiving unit. The other factor in the rule, aj, is just the activation of the sending unit. As long as ε is small, this rule will tend to move connection weights toward values that bring the internal input to the unit into alignment with the external input. For example, if the internal input is smaller than the external input, and the activation of the sending unit is positive, the weight between them will be increased; this will increase the internal input, reducing *delta*.

McClelland and Rumelhart (1985) used this model to illustrate several key points about auto-associators. One is that they can act as content-addressable memories. For example, different subsets of the units might represent different aspects of a particular object. Some might represent the sight of a rose, some its aroma, some its name, some what it feels like to touch, and so on. Presentation of some aspects of the pattern will tend to lead to reconstruction of other aspects, filling them in on the other units. Another is that the network tends to clean up noisy versions of patterns on which it has been trained. Any pattern that is similar to one of the stored patterns will tend to be modified so that it is more like the stored pattern. In networks of this type, the stored patterns are often called "attractors" because patterns that are near them tend to change over time, in the direction of becoming more similar to the stored patterns. Each attractor has a basin of attraction around it, representing the set of patterns that gets attracted to it. Both pattern completion and pattern cleanup occur in settling toward an attractor state.

Because the learning that occurs for one pattern affects the same set of weights that are used by all other patterns, there is an automatic tendency for similarity-based generalization and blending of similar patterns to occur. Thus models of this type address blend errors and prototype effects seen in memory experiments. Some authors have suggested that blend errors and prototype effects are best accounted for by models in which each study item is stored separately. In models of this type (e.g., Hintzman, 1988), presentation of a test item results in activation of all similar studied items, and the resulting activations are then summed together (this idea can be implemented in localist connectionist models such as McClelland, 1981). This makes it possible to account for the fact that, in addition to an advantage for the prototype of a category over particular studied items, there is generally also an advantage for studied items over other nonprototypical items in the category. However, McClelland and Rumelhart (1985) showed that both prototype and item effects occurred in their distributed model as well. Indeed, the connection weights in superpositional, distributed models can capture proto-

type and instance characteristics of several different categories at the same time, thus obviating the need to suppose that memory contains separate stored representations of every item ever experienced.

The ideas and models described above arose within the context of models that made use of either a coincident activation learning rule or an error-correcting learning rule. In the application of these ideas to models of memory and related phenomena, modelers stipulated the activations of input and output units, and used these rules to adjust the strengths of direct connections from the input units to the output units (in auto-associator models, the input and output units may be the same units). It was known from the beginning of this work that such models faced a serious computational limitation. In essence, when two input patterns activate many units in common, the patterns will tend to produce the same output, since both patterns will be using the same set of connection weights. If the outputs must be different, the few units that differ between the patterns will be left to do all of the work. If the units in question must do different work in different cases, a situation can arise in which there is no set of weights that will address all cases at once.

One way to solve this problem is to provide a mechanism for expanding the representation of an input, by pre-specifying units to represent possible conjunctions or combinations of inputs. This approach has been used successfully in some models (e.g., Gluck & Bower, 1988). However, the set of all possible conjunctions grows exponentially with the number of inputs, so that even with only 40 individual elements, the number of conjunctions exceeds the number of neurons in the brain (thought to be about 10^{11}, or 100 billion). This problem can be addressed to a degree. In one sort of solution (e.g., O'Reilly & McClelland, 1994), each of a number of conjunctive units receives inputs from a random subset of the input units, and the conjunctive units with the largest number of active inputs are chosen to represent the current input. Another approach is to create a large number of initially uncommitted units with random weights to all elements of the input. When an input is presented, the unit that it activates most strongly based on the random initial weights is selected, and its weights are tuned to match the input. This approach was taken by Grossberg (1976, 1978) and Kruschke (1992, 1996), both of whom have applied it to aspects of human memory. A very different approach to the problem arose from the use of a sophisticated version of the error-correcting learning rule to discover useful internal representations of inputs. We will consider this approach in the context of a model of the representation and use of knowledge in semantic memory.

Connectionist vs. Symbolic Models of Semantic Memory

The classical, symbolic approach to semantic memory was proposed by Quillian (1968). It involved supposing that such information consists of a set of propositions organized into a taxonomic hierarchy. In the example shown in figure 36.3, knowledge of various kinds of plants and animals is represented hierarchically under the common superordinate "Living Thing." All propositions include a concept name, one of four relations, and another concept or property, as in the examples "canary is a bird," "canary is yellow," "bird has feathers," and "animal can move." Note that the ISA propositions encode the backbone structure of the hierarchy. According to Quillian's proposal, propositions that are true of all the concepts within a certain branch of the tree would be stored at the top of the branch; thus, since all animals can move, that proposition is stored with *animal*, but since only the birds can fly, that proposition is stored with *bird*. The result is that information is stored economically, and many inferences can easily be derived from the stored information. Thus if we simply add the proposition "sparrow is a bird," we can then infer that it can fly, that it has feathers, that it can move, etc., by following ISA propositions up the tree and reading out what is stored there.

Quillian's proposals sparked a great deal of interest in the 1970s. However, the approach is somewhat brittle in the face of exceptions and doesn't provide a very natural way of dealing with typicality effects (Rosch, 1975). Also, experiments showed that general properties did not take longer to access than specific ones, and Rips, Shoben, and Smith (1973) showed that sometimes general category membership could be verified more easily than more specific category membership. An alternative, connectionist approach that captures many of the desirable properties of Quillian's proposal and does not suffer from these difficulties was provided by Rumelhart

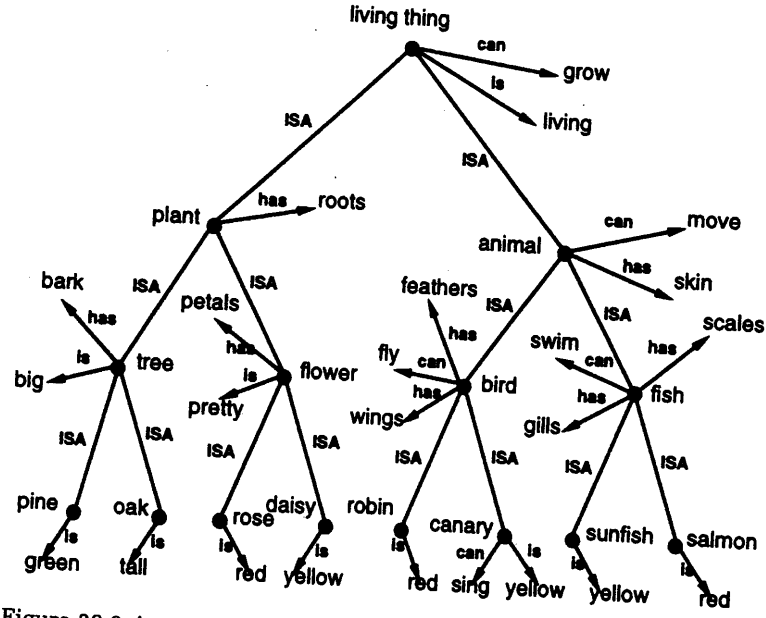

Figure 36.3 A semantic network of the type formerly used in models of the organization of knowledge in memory. All of the propositions used in training the network are based on the information actually encoded in this figure. For example, the network indicates that living things can grow; that a tree is a plant; and that a plant is a living thing. Therefore it follows that a tree can grow. All of these propositions are contained in the training set. *Note*: Redrawn with alterations from D. E. Rumelhart and P. M. Todd (1993), Learning and connectionist representations (figure 1.8, p. 14), in *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, edited by D. E. Meyer and S. Kornblum, Cambridge, MA: MIT Press. Copyright 1993 by MIT Press. Permission pending.

(1990; Rumelhart & Todd, 1993). They used the sophisticated error-correcting learning algorithm mentioned previously to train a network on the propositions stored in Quillian's hierarchical network, and showed that after training it too exhibited parsimonious representations and allowed inferences to be made from newly learned concepts. This model propagated activity in only one direction, and so it is more like the pattern associator model of the rose-steak example above. An auto-associative model with similar properties has been implemented by (O'Reilly, 1996), but we consider the Rumelhart version for simplicity.

Rumelhart's (1990) network is shown in figure 36.4. Connections run from a pool of concept units and a pool of relation units to output units for the concepts and for each of the different types of properties (*is*, *has*, and *can* properties). Note there are two other pools of internal or "hidden" units, between the input and the output. One pool, called *concept representation* units, sits between the concept units and the other pool, called general hidden units, which receives input from the concept representation units and the relation units. This network is initialized with small random weights on all the connections. Testing of the network takes place by activating a concept input unit and a relation unit, with all other input units off, and propagating activation forward through the network. Thus we can test the network's knowledge of what a canary can do by activating *canary* and *can* on the input. Activation propagates forward: net inputs are calculated as in equation 2, with the activation of the unit a monotonic, S-shaped function of the net input.

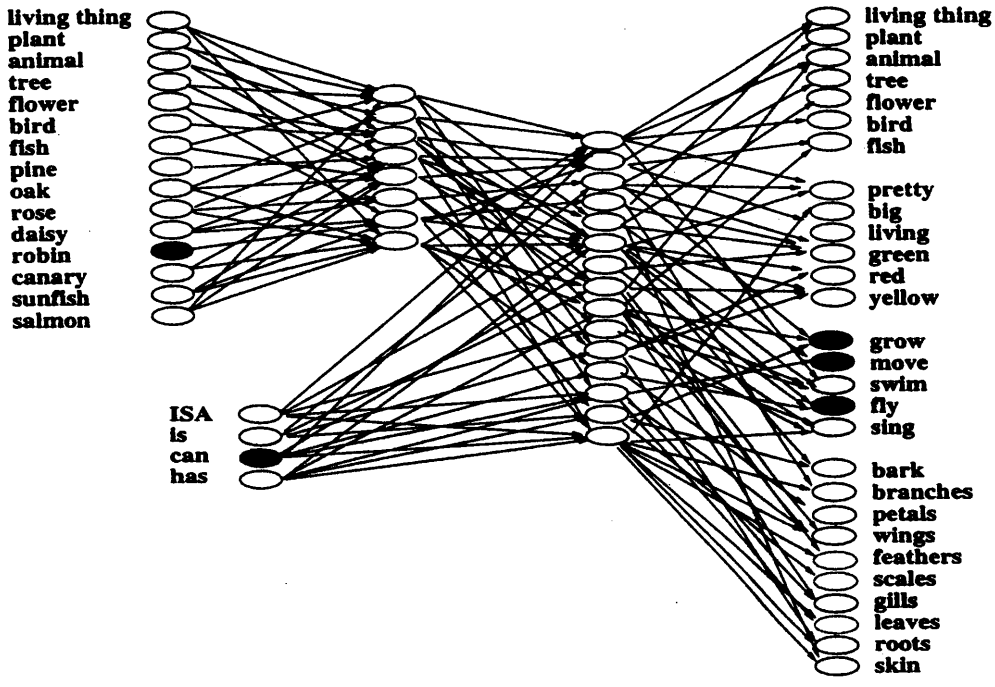

Figure 36.4 A depiction of the connectionist network used by Rumelhart to learn propositions about the concepts shown in figure 36.3. The entire set of units used in the actual network is shown. Inputs are presented on the left, and activation propagates from left to right. Where connections are indicated, every unit in the pool on the left (sending) side projects to every unit in the right (receiving) side. An input consists of a concept-relation pair; the input *robin* can is illustrated here by darkening the active input units. The network is trained to turn on all those output units that represent correct completions of the input pattern. In this case, the correct units to activate are *grow*, *move* and *fly*; the units for these outputs are darkened as well. Subsequent analysis focuses on the concept representation units, the group of eight units to the right of the concept input units. Reprinted from McClelland et al. (1995), figure 5, p. 430, based on the network depicted in Rumelhart and Todd (1993), figure 1.9, page 15.

Training the Network with Back Propagation

At first, owing to the small random weights, the activations of the hidden and output units take on values that hover around the neutral value of 0.5. The network must learn the content of the domain through training. The training procedure, called back propagation, was developed independently by Rumelhart, Hinton, and Williams (1986) and other groups. Training occurs in a series of epochs; in each epoch, each input pattern is presented (i.e., each combination of a concept and a relation), and the output is generated as described above. The output is compared to the correct, target output, and weights are adjusted according to the back propagation learning algorithm, which adjusts each weight a very small amount to reduce the difference between the desired and the obtained activation. For the weights projecting from the general hidden layer to the output units, a procedure very similar to the error-correcting learning rule presented in equation 3 is used. For each unit, a quantity called $delta_k$, representing the extent to which changing the net input to the unit would reduce the network's overall error, is calculated. For output units, this quantity is equal to the difference between the target value t_k and the obtained activation a_k, times a

scaling factor depending on the activation of the unit

$$\delta_k = (t_k - a_k)s(a_k). \quad (4)$$

The important innovation in the learning procedure is that it provides a way to adjust the connection weights coming into hidden units from units lower down in the processing stream. Essentially, a delta term is assigned to each such unit by considering how a change in its input would affect the error at each of the units to which it projects, and then adding up these separate effects (with a scaling factor as before):

$$\delta_i = \Sigma_k w_{ki} \delta_k s(a_i). \quad (5)$$

Here Σ_k runs over all the units to which hidden unit i has a forward connection; w_{ki} is the weight on one such connection, and $delta_k$ is the *delta* term for the unit at the end of that connection. Thus the *delta* of each hidden unit is essentially a weighted sum of the *deltas* of all the units to which that hidden unit projects. The algorithm is called back propagation because each hidden unit's *delta* is calculated by "propagating" the delta terms associated with downstream units "backward" across the connection weights. Once *deltas* have been calculated for all of the hidden units (and this process must proceed backward by layers from the output units), all the forward-going weights to each hidden unit from all the units below it can now be adjusted according to equation 5, where i is understood to indicate the unit at the forward or receiving end of the connection and j indicates the unit at the sending end.

Back-propagation has often been criticized because, taken literally as a procedure for training connections in the brain, it appears biologically implausible. While this criticism may have some merit, it is far from clear exactly what takes place biologically. Furthermore, back propagation and other similar learning algorithms open up vast new possibilities for understanding human cognition and its development as an adaptive learning process. This is illustrated by an analysis of the outcome and time course of learning in Rumelhart's semantic network.

Cognitive and Developmental Implications

The first thing to notice about the network is that it learns by a process that might be called progressive differentiation. This process affects both the output of the network and also its internal representations of concepts, as can be seen in figure 36.5 from a subsequent study of this network by McClelland, McNaughton, and O'Reilly (1995). The figure shows the patterns of activation assigned by the network to each of the lowest-level concepts in the hierarchy, at three stages of learning: (1) very early in learning, where the representations are determined primarily by the initial random values of the connection weights from the concept input to the concept representation units; (2) midway through learning; and (3) at the end of learning. For each concept at each point in learning, the activations of the eight concept representation units that are produced by activation of the corresponding concept unit are shown. Early in learning, the patterns or activation the network assigns to the different concepts are all very similar, with none of the units either strongly on or strongly off in any of the patterns. Midway through learning, the network has differentiated the plants from the animals, but within the plants and the animals, there are only very slight differences. At the end of training, the network has differentiated the birds from the fish and the flowers from the trees, and there are also subtle but important differences in the representations that the network assigns to the individual birds, fish, trees, and flowers. This process appears to be consistent with a corresponding process of progressive differentiation seen in child development (see McClelland et al., 1995, for discussion).

The second thing to note about the network is that it uses a parsimonious approach to representation of semantic information. By assigning very similar representations to concepts for which very similar sets of propositions are true, it can use the same set of connection weights forward of these representations to answer these questions. This use of similar representations for concepts that share propositions also allows the network to generalize what it has learned about one concept to other similar concepts. Indeed, what is common across similar concepts is more robustly represented than what is idiosyncratic. This allows the network to account naturally for the pattern of semantic loss seen in patients with a progressive deterioration of semantic memory (Warrington, 1975; Hodges, Graham, & Patterson, 1995). These patients show deterioration of knowledge of details of con-

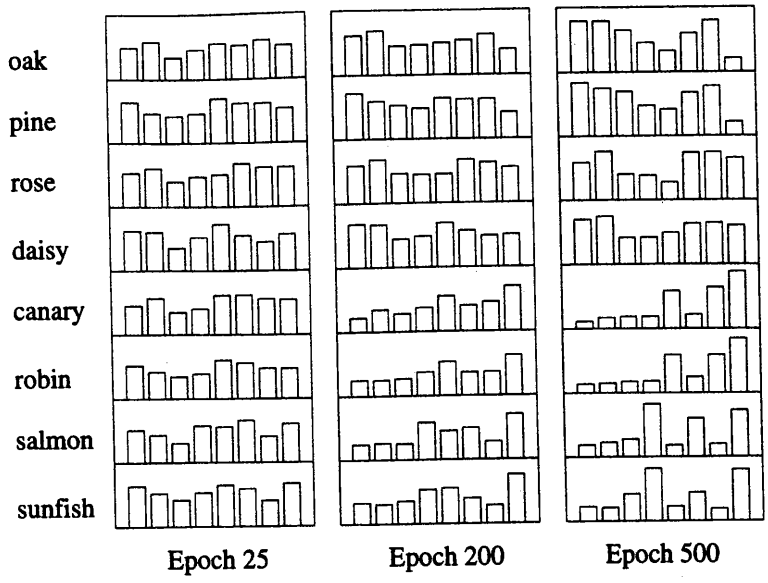

Figure 36.5 Representations discovered in our replication of Rumelhart's learning experiment, using the network shown in figure 36.4. The figure presents a vertical bar indicating the activation of each of the eight concept representation units produced by activating the input unit for each of the eight specific concepts. The height of each vertical bar indicates the activation of the corresponding unit on a scale from 0 to 1. One can see that initially all the concepts have fairly similar representations. After 200 epochs, there is a clear differentiation of the representations of the plants and animals, but the trees and flowers are still quite similar as are the birds and the fish. After 500 epochs, the further differentiation of the plants into trees and flowers and of the animals into fish and birds is apparent. Reprinted from McClelland et al. (1995), figure 36.6, p. 431.

cepts, while still retaining their general characteristics. The same pattern is observed when progressive damage is simulated by the addition of increasing amounts of noise into the representations used in the network (McClelland & Rogers, 1997).

Catastrophic Interference and Complementary Systems in Memory

Connectionist learning procedures like back propagation that train hidden units offer new ways of thinking about semantic memory quite different from those provided by traditional semantic network accounts that trace their heritage back to Quillian (1968). It was, therefore, something of a disappointment when McCloskey and Cohen (1989) applied back propagation to a classic memory paradigm—paired associate learning—and discovered that the algorithm appeared to suffer from a problem that they labeled "catastrophic interference." They trained a connectionist network on a list of eight paired-associate patterns, then looked to see the effect on these learned associations of subsequent training on a second list of eight new associations involving the same input patterns but different responses. The experiment is analogous to many classical memory experiments of retroactive interference. Human subjects do show some interference from learning a second list of associations on recall of a first list, but in the network the interference was far more profound. In fact, before the network correctly produced any of the new associations its ability to reproduce all of the old associations cor-

rectly was completely wiped out. The only way to avoid the catastrophic interference problem was to interleave learning of one set of associations with learning of the other set. In this way, the learning procedure could find a set of connection weights consistent with correct performance on both sets of associations.

This finding of catastrophic interference might seem to suggest that the back propagation model might be fatally flawed as a model of human learning and memory. However, another approach—the one taken by McClelland et al. (1995)—was to suggest that the catastrophic interference effect found by McCloskey and Cohen (1989) might be avoided in the human brain by using a back-propagation-like learning system for semantic and procedural learning, while using a different kind of learning system for the initial learning of arbitrary associations. They suggested that the areas of the neocortex of the brain in which semantic knowledge is stored may learn very slowly, so that learning of any one item can be interleaved with ongoing learning of other items. As we saw in the consideration of the learning of semantic information about plants and animals, this results in discovery of efficient representations that support generalization and robustly represent information that is shared by many things. They further suggested that other areas of the brain—particularly the medial temporal lobes, where the hippocampus and related brain structures are found—may be specialized for the rapid storage of arbitrary new information, such as paired associates. Considerable neuropsychological evidence supports the idea that the medial temporal regions of the brain plays this role; for example, individuals with extensive damage to this region are profoundly deficient in the initial acquisition of arbitrary new factual information, including episodic memory (memory for the particular contents of particular events and experiences) and paired-associate learning. What McClelland et al. (1995) proposed is that the learning system in the medial temporal lobes makes use of a conjunctive scheme like that of O'Reilly and McClelland (1994) to minimize overlap of distinct memories. Such coding schemes have been used by several investigators to model paired-associate learning without catastrophic interference (French, 1991, 1992; Sloman & Rumelhart, 1992; Kortge, 1993). Evidence from single-neuron recording studies in animals supports the idea that the hippocampus uses very sparse, conjunctive representations (e.g., Barnes, McNaughton, Mizumori, Leonard, & Lin, 1990).

Given that the catastrophic interference problem can be avoided by the use of sparse conjunctive coding, why would this approach not be used for semantic as well as episodic memory? The answer to this is key to understanding why there must be different types of learning systems in the brain. While sparse random conjunctive coding allows rapid learning of new memories, it does so in a very simple way. It assigns a distinct representation, minimizing overlap with other memories. Back propagation, on the other hand, and other learning algorithms that exploit gradual, interleaved learning, find patterns of connection weights that capture the overall structure of entire ensembles of events and experiences, and that assign representations to individual concepts that capture the position of the concept in that structure. In this regard it is interesting that the input to the medial temporal lobe systems appears to come from the neocortical regions responsible for semantic and other relatively abstract forms of representation. The arrangement suggests gradual, interleaved learning may be used to develop the representations that are then made available for storage, so that what is stored in episodic memory is not a raw copy of sensory input, but the results of sophisticated representational processes acquired throughout development in the neocortical cognitive system.

In summary, this section has reviewed different kinds of connectionist approaches to overcoming the limitations of networks containing a single layer of modifiable weights. Two kinds of solutions to these problems were considered. One of these involved some variant of a sparse random, conjunctive coding scheme, and the other involved the use of interleaved learning via back propagation to discover the structure of a domain and assign appropriate representations to concepts within it. There are strengths and weaknesses of each approach, but it appears that the brain solves the dilemma posed by this situation by making use of both solutions. Thus research on connectionist models provides a basis for understanding what might otherwise appear to be a relatively arbitrary aspect of brain organization—namely, that there appear to be separate systems in the brain subserving the initial acquisition of arbitrary information on the one hand and the systematic representation of knowledge in semantic memory on the other.

Current Directions in Connectionist Models of Memory

Attentional and Strategic Aspects of Memory

Several connectionist models now take note of the fact that encoding and retrieval of information is subject to strategic manipulation (see Cohen & O'Reilly, 1996, for a general consideration of such effects). Kruschke (1992) showed how a connectionist network could provide a mechanism implementing the distribution of attention to dimensions of input stimuli. In this earlier work and in much of his later work these ideas have been applied to learning in categorization tasks, but Dennis and Kruschke (1998) have now shown that some findings that justify the claim that attention to stimulus cues adjusts adaptively in categorization experiments are also obtained in cued-recall memory experiments. In particular, subjects appear to reduce attention to cues that they discover are ambiguous. Kruschke (1996) provides a connectionist model that captures these effects in categorization experiments and the model is extended to attention effects in memory in Dennis and Kruschke (1998). The authors suggest that such attentional effects influence the pattern of activation that arises in the neocortex when an input is experienced, affecting what is available for storage in the medial temporal lobe memory system and ultimately what is available for integration into the knowledge base stored in connections among neurons in the neocortex.

Relation of Connectionist Models to Bayesian and Other Optimal Approaches

A recent issue in memory research concerns whether it is useful to construe human memory as optimal. This question has been posed forcefully by Anderson (1990; Anderson & Milson, 1989), who has suggested that, indeed, several aspects of memory can be construed to reflect the use of optimal policies for storage, retrieval, and so on. Interestingly, there are several ways in which connectionist models may be seen as optimal or quasi-optimal (McClelland, 1998). Perhaps most fundamentally, there is a very strong connection between connectionist learning procedures and procedures for deriving optimal estimates of parameters for complex nonlinear estimation procedures (White, 1989; MacKay, 1992). Individual connectionist units and indeed whole connectionist networks can be construed as optimal Bayesian estimators of conditional probabilities of hypotheses given evidence, and several connectionist learning rules can be viewed as procedures for deriving estimates of necessary probability relations between hypotheses and evidence in connection weights. Also, a Bayesian, probability-estimation approach has recently been applied to modeling recognition memory (Shiffrin & Steyvers, 1997; McClelland & Chappell, 1998), and McClelland and Chappell (1998) have indicated how such models might be implemented in connectionist networks.

Hebbian vs. Error-Correcting Learning Rules

The final current research direction we will consider in this article tends to undercut the idea that human learning and memory are in any general sense optimal, however. This is the idea that, perhaps, certain aspects of human learning may reflect Hebbian as opposed to error-correcting synaptic adjustment rules. Hebb's original idea was that synaptic modification works to strengthen the connection from one neuron to another, when the first appears to persistently or repeatedly take part in firing the second (the coincident activation learning rule is one example of a Hebbian-type learning rule). This rule may sometimes be counterproductive: If a stimulus activates a set of input units, and these in turn activate a pattern on some other set of units, then Hebbian learning will tend to reinforce whatever that pattern turns out to be, whether or not it is the desired response in a memory experiment. This may have the effect of reinforcing preexisting, incorrect response tendencies, and thus of actually impeding rather than enhancing progress in memory experiments.

If this is correct, the human memory system can behave far from optimally in many cases. Indeed, the idea that the mechanisms of synaptic plasticity operate according to the Hebbian principle leads to the observation that these mechanisms might be instrumental in the maintenance of maladaptive and highly nonoptimal behavior in many cases. This serves to reinforce the observation that rationality or optimality is generally conditional on the accuracy of certain (explicit or implicit) assumptions. Error-correcting learning with

an exponentially decreasing learning rate can be optimal when learning in an environment when the training examples are sampled from a distribution that remains invariant over time, but is highly nonoptimal if, after considerable experience, the distribution of training examples changes. Hebbian learning may be optimal or approximately so, as long as conditions are arranged so that the activations produced by an input are predominantly desirable, but can be completely counterproductive in cases where activations produced by inputs are not the ones that are desired.

Conclusion

Connectionist models have been developed to capture what their originator's suggest may be fundamental aspects of human learning and memory systems. In most cases, these models bear strong similarities to nonconnectionist models. This observation is consistent with the idea that there may be many different frameworks within which some ideas can be captured. The connectionist framework has been fertile in allowing a range of different principles to be explored—so fertile that the framework itself is often criticized as unhelpfully general or open-ended (Massaro, 1988). While this open-endedness does have its downsides, it also has several benefits, one of which is that it has allowed the exploration of a wide range of different ideas about the nature of learning and memory. Another benefit is the naturalness with which it may be applied to addressing neuropsychological phenomena, and the ease with which it can be used to adopt specific proposals from neuroscience, including the idea that synaptic modification follows the principles of Hebbian learning, or that sparse, conjunctive representations appear to be used in the hippocampus. The approach appears to provide a useful complement to other more constrained models, many of which appear to be highly applicable to data obtained within certain classes of paradigms, but which often have little to say outside their range of applicability and which may not immediately suggest ways of incorporating findings from neuroscience. It thus appears that connectionist models play a useful role in our efforts to understand the nature of human memory, complementing other approaches.

References

Anderson, J. A. (1973). A theory for the recognition of items from short memorized lists. *Psychological Review, 80,* 417–438.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84,* 413–451.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96,* 703–719.

Arbib, M. A. (1995). *The handbook of brain theory and neural networks.* Cambridge, MA: MIT Press.

Barnes, C. A., McNaughton, B. L., Mizumori, S. J. Y., Leonard, B. W., & Lin, L.-H. (1990). Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Progress in Brain Research, 83,* 287–300.

Cohen, J. D., & O'Reilly, R. C. (1996). A preliminary theory of the interactions between prefrontal cortex and hippocampus that contribute to planning and prospective memory. In M. Brandimonte, G. O. Einstein, & M. A. McDaniel (Eds.), *Prospective memory: Theory and applications.* Mahwah, NJ: Erlbaum.

Dennis, S., & Kruschke, J. K. (1998). Shifting attention in cued recall. *Australian Journal of Psychology, 50,* 131–138.

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

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

Gluck, M. A., & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language, 27,* 166–195.

Grossberg, S. (1976). On the development of feature detectors in the visual cortex with applications to learning and reaction-diffusion systems. *Biologcial Cybernetics, 21,* 145–159.

Grossberg, S. (1978). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans.

Progress in Theoretical Biology, 5, 233–374.

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (pp. 161–187). Hillsdale, NJ: Erlbaum.

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

Hodges, J. R., Graham, N., & Patterson, K. (1995). Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory, 3*, 463–495.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review, 96*, 208–233.

Knapp, A., & Anderson, J. A. (1984). A signal averaging model for concept formation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 617–637.

Kortge, C. A. (1993). Episodic memory in connectionist networks. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 764–771). Hillsdale, NJ: Erlbaum.

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

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition, 22*, 3–26.

Lashley, K. S. (1950). In search of the engram. In *Society of Experimental Biology Symposium* (pp. 478–505). London, UK: Cambridge University Press.

Loftus, E. (1991). Made in memory: Distortions of recollection after misleading information. In G. Bower (Ed.), *Psychology of learning and motivation* (Vol. 27; pp. 187–215). New York: Academic Press.

MacKay, D. J. (1992). Information-based objective functions for active data selection. *Neural Computation, 4*, 590–604.

Massaro, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory and Language, 27*, 213–234.

McClelland, J. L. (1981). Retrieving general and specific information from stored knowledge of specifics. In *Proceedings of the Third Annual Conference of the Cognitive Science Society* (pp. 170–172). Berkeley, CA.

McClelland, J. (1998). Connectionist models and Bayesian inference. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 21–53). Oxford, UK: Oxford University Press.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*, 724–760.

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.

McClelland, J. L., & Rogers, T. T. (1997). A PDP account of basic-level category effects and semantic demantia (Abstract 133). *Abstracts of the Psychonomic Society, 2*, 14.

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.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1; chap. 1, pp. 3–44). Cambridge, MA: MIT Press.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 24; pp. 109–165). New York: Academic Press.

McNaughton, B. L. (1993). The mechanism of expression of long-term enhancement of hippocampal synapses: Current issues and theoretical implications. *Annual Review of Physiology, 55*, 375–396.

Metcalfe, J. (1990). Composite holographic associative recall model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General, 119*, 145–160.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*, 609–626.

Nystrom, L. E., & McClelland, J. L. (1992). Trace synthesis in cued recall. *Journal of Memory & Language, 31*, 591–614.

O'Reilly, R. (1996). *The leabra model of neural interactions and learning in the neocor-*

tex. PhD thesis, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus, 4*, 661–682.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353–363.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing* (pp. 216–270). Cambridge, MA: MIT Press.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12*, 1–20.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192–233.

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.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1; pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 3–30). Cambridge, MA: MIT Press.

Shiffrin, R. M., & Steyvers, M. (1997). A model of recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145–166.

Sloman, S. A., & Rumelhart, D. E. (1992). Reducing interference in distributed memories through episodic gating. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *Essays in honor of W. K. Estes* (Vol. 1; pp. 227–248). Hillsdale, NJ: Erlbaum.

Warrington, E. (1975). Selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology, 27*, 635–657.

White, H. (1989). Learning in artificial neural networks: A statistical perspective. *Neural Computation, 1*, 425–464.

Willshaw, D. (1981). Holography, associative memory, and inductive generalization. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (Chap. 3; pp. 83–104). Hillsdale, NJ: Erlbaum.