

Connectionist Approaches*

B. J. MacLennan
Computer Science Department
University of Tennessee, Knoxville
maclennan@cs.utk.edu

September 5, 2001

Abstract

Connectionist approaches to cognitive modeling make use of large networks of simple computational units, which communicate by means of simple quantitative signals. Higher-level information processing emerges from the massively-parallel interaction of these units by means of their connections, and a network may adapt its behavior by means of local changes in the strength of the connections. Connectionist approaches are related to neural networks and provide a distinct alternative to cognitive models inspired by the digital computer.

1 Definitions

To facilitate the following discussion, it will be helpful to define some terms. A typical connectionist network comprises a (potentially large) number of simple processing *units*. The units are often called (artificial) neurons, but that terminology begs the question of their relation to biological neurons, so it will be avoided here (Sect. 5.4). In the most common case, the units form a weighted sum of their (quantitative) inputs and pass the result through a simple, nonlinear *activation function*, which limits the range of possible outputs. The resulting value is considered the *activity* of the unit, which may be transmitted to other units (through outgoing connections). In some cases the activity of a unit is a combination of its inputs and previous activity, which provides a kind of “short-term memory” residing in the collective activities of the units.

The weighted sum results from the fact that each connection in the network has an associated *weight* (analogous to synaptic efficacy in biological neural networks), which multiplies the quantity transmitted by that connection. Positive weights correspond to excitatory connections and negative weights to inhibitory;

*To appear in *International Encyclopedia of the Social and Behavioral Sciences* (26 vols.), ed. by Neil J. Smelser & Paul B. Baltes, Pergamon, in press. This is a preprint, and not to be reproduced without permission.

zero-valued weights correspond to the absence of a connection. Mathematically, connection weights are often treated as a *weight matrix* \mathbf{W} , with element W_{ij} being the weight of the connection to unit i from unit j . Learning and adaptation take place by modification of the weights according to some *learning algorithm* (Sect. 3); thus the connections constitute the network's "long-term memory." "Connectionism" derives its name from the fact that knowledge resides in the patterns and weights of the connections.

Many connectionist networks are organized into *layers*, analogous to functional areas in the brain; information usually moves in lockstep from layer to layer. Although many networks are *feed-forward*, that is, the information moves through successive layers from input to output, other networks are *recurrent*, which means that there may be *feedback* connections from a layer to itself or to earlier layers. Recurrent networks are able to recognize and process *temporally-extended patterns*, that is, sequences of related inputs.

It must be stressed that there are exceptions to all of the preceding general statements about connectionist networks, and "connectionist approaches" are best viewed as forming a Wittgensteinian "family resemblance."

2 History

A short history of connectionist approaches will be presented, first in the narrower context of cognitive science and artificial intelligence, then in the broader context of epistemology, linguistics and the philosophy of mind. Although the size of this article does not permit detailed citations of the literature, many of the seminal articles are collected in Anderson & Rosenfeld (1988) and Anderson, Pellionisz & Rosenfeld (1990). *See also:* Cognitive Psychology, History; Cognitive Science, History.

2.1 Narrower History

According to the second edition of the *Oxford English Dictionary*, the term "connectionism" was first used by E. L. Thorndike in his *Fundamentals of Learning* (1932) to refer to the reduction of mental processes to the connections between stimuli and responses, that is, to a form of *associationism*, and so connectionist theories have been set in opposition to cognitive theories. The term is used somewhat differently now (so that "connectionist cognitive science" is not an oxymoron), but retains some similarities to associationism. However, to understand the relation it is better to look at connectionism from the perspective of neural network models of cognition.

In the early 1940s W. S. McCulloch and W. Pitts investigated the computation of logical functions by simple neuron-like elements; in effect they showed that these elements could compute logical "and," "or," "not," and so forth. In his *Organization of Behavior* (1949) D. O. Hebb suggested that learning takes place by the formation of *cell assemblies*, and that this occurs through the strengthening of connections between simultaneously active neurons in neural networks,

which are initially randomly connected. His description of this process inspired one of the simplest connectionist learning rules (Sect. 3.1).

In the late 1950s F. Rosenblatt began to investigate the application of simple neuron models called *perceptrons* to perceptual problems such as classifying printed letters (see *Perceptrons*). He developed a learning algorithm for simple (single-layer) perceptron networks, which iteratively adjusted the connection weights whenever the network made a mistake. He proved that if the network were capable of solving the problem at all, then the algorithm would eventually find the connection weights to solve it. However, there are many problems that a single-layer network cannot solve, and Rosenblatt never succeeded in finding a multilayer learning algorithm.

A key event in the history of connectionism was the publication of M. Minsky and S. Papert's *Perceptrons* (1969), which demonstrated limitations of simple perceptron networks. Specifically, they proved that single-layer perceptron nets could discriminate only those categories that are *linearly separable* (see *Linear Algebra for Neural Networks; Perceptrons*). Their proof did not apply to multilayer nets (for which there was still no learning algorithm), but they suggested that similar limitations would be found for these too. Nevertheless, their book was widely interpreted as showing the impotence of neural networks in general, and is commonly blamed for discouraging research in the field for a decade. (The extent to which it actually did so is, perhaps, a topic for historians of science.)

It will be worthwhile to comment on the role of holography in the development of connectionist approaches. As early as 1929 K. S. Lashley had conducted experiments suggesting that individual memory traces were not *localized* in any one place, and that degradation of memory was proportional to the amount of cortical mass destroyed, thus implying that individual traces were *distributed* over large areas of cortex. In well-known 1950 paper, he despaired of ever understanding how such a *nonlocal* memory could operate. The principles of holography had been described by D. Gabor in 1949, but it was not until the advent of optical holography in the early 1960s that it began to be seen as a solution to Lashley's dilemma. Although an analogy between holography and memory had been suggested as early as 1963 by P. J. van Heerden, the "holographic hypothesis" has been developed most extensively by K. H. Pribram and his colleagues since 1966 (see, e.g., Anderson, Pellionisz & Rosenfeld 1990, ch. 7). In the late 1960s and early 1970s holographic and holography-inspired models of associative memory were also investigated by H. C. Longuet-Higgins, D. J. Willshaw and others (Hinton & Anderson 1989). Some connectionists were influenced by the critiques of traditional, rule-based AI by the phenomenologist philosopher H. L. Dreyfus (*What Computers Can't Do*, 1972). Although he stressed the limitations of rule-based systems, he also suggested that some of these limitations would not apply to analog systems operating on principles similar to holography. *See also:* Distributed Cognition; Memory Trace, Nature of; Models of Neural Basis of Learning and Memory.

Although a number of investigators (including J. A. Anderson, S. Grossberg and T. Kohonen) continued connectionist research through the 1970s, the

field was rejuvenated by the work of D. E. Rumelhart, J. L. McClelland, and other members of the “PDP (Parallel Distributed Processing) Working Group,” many of whose publications were collected in a widely-read two-volume set (Rumelhart, McClelland & the PDP Research Group 1986). The credibility of connectionist approaches was also enhanced by J. Hopfield’s publication in 1982 of a simple recurrent net capable of associative memory and pattern completion.

2.2 Broader History

Although connectionism can be viewed as an approach to knowledge representation and inference of relevance only to cognitive science, in fact it has much broader implications, for it challenges assumptions about knowledge that have been largely unquestioned since ancient Greek philosophy. Already in the philosophies of Socrates, Plato and Aristotle there is a preference for knowledge expressed as logical relations among discrete, language-like structures, and for a view of cognition as mechanized deduction. These ideas influenced many later philosophers, including Hobbes (who equated thinking with computation), Leibniz (who experimented with formalized systems of knowledge representation and mechanical deduction), and Boole (who invented mathematical logic). These ideas were also influential in the development of logical positivism, which dominated the philosophy of science in the first half of the twentieth century. The idea persisted in the assumption that there must be a “language of thought,” because no alternative is imaginable, and it is “the only game in town.” Similarly, most research in artificial intelligence (AI) took for granted that intelligence resides in the structures of a *knowledge representation language* and in deduction-like formal rules for their manipulation. Throughout the 1970s (the “connectionist dark ages”), AI researchers concentrated their attention on *expert systems*, which depended on expertise represented symbolically (see *Expert Systems*). Disappointment with the performance of these systems was one of motivations for the connectionist renaissance.

In summary, the Western tradition (with some exceptions) has displayed a kind of “linguistic chauvinism,” which presumes that all knowledge and cognition can be expressed in language-like structures. Knowledge is expressed at a *symbolic* level, that is, in terms of atomic (indivisible), word-level categories related by sentence-like logical structures. On the other hand, most connectionist approaches represent knowledge at a *subsymbiotic* level, that is, in terms of minute, quantitative features related by low-level, often statistical, connections. In other words, knowledge is more akin to an *image* than to a *sentence* (see *Imagery vs. Propositional Reasoning; Mental Imagery, Psychology of*). Therefore, some of connectionism’s advocates see it as a fundamentally new view of knowledge and cognition, which is leading to a paradigm shift in cognitive science and philosophy and is engendering a “new AI.” Specific innovations of the connectionist approach are discussed in Sect. 5.1. *See also:* Artificial Intelligence: Connectionist and Symbolic Approaches; Knowledge Representations, Theory of; Production Systems, in Cognitive Psychology; Schemas, Frames and Scripts,

in Cognitive Psychology; Symbolic Approaches, in Cognitive Science.

3 Mechanisms of Adaptation and Learning

Virtually all connectionist approaches incorporate adaptive mechanisms or *learning algorithms*, which allow the network to improve its performance; here a few will be discussed briefly.

3.1 Correlational Learning

Correlational learning (“Hebb’s rule”), the simplest connectionist learning algorithm, takes its inspiration from Hebb’s hypothesis that the simultaneous activity of two neurons strengthens the connection between them. It makes a change in connection weight proportional to the product of activities of the units it connects. Thus, the change in the weight W_{ij} of the connection to unit i from unit j is proportional to $y_i x_j$, where y_i is the activity of unit i and x_j is the activity of j . This learning rule can be viewed as a highly simplified model of *long-term potentiation*. See also: Long-term Potentiation and Depression (Cortex); Models of Neural Basis of Learning and Memory; Regulation of Synaptic Efficacy.

The effect of this rule is that the weight becomes a correlation coefficient between the activities of the units it connects. That is, the connection will become stronger (more positive) to the extent that the units are simultaneously positive or simultaneously negative. The connection will become more inhibitory (more negative) to the extent that one unit is positive while the other is negative. If there is, on average, no systematic relation between the activities of the two units, then the weight will tend toward zero, effectively disconnecting the units.

The correctional learning rule is the basis of a simple associative memory known as a *linear associator*. In this connectionist network, there are two layers of linear units, an input layer and an output layer; each input unit is connected to every output unit, so that the output is a linear function of the input, $\mathbf{y} = \mathbf{W}\mathbf{x}$. A series of pairs of pattern vectors $(\mathbf{y}_1, \mathbf{x}_1), \dots, (\mathbf{y}_p, \mathbf{x}_p)$ may be presented to the input and output layers of such a network and the weights adjusted according to the learning rule. The goal of the linear associator is that the network associate each \mathbf{x}_k with the corresponding \mathbf{y}_k . In fact it can be shown that if the set of input patterns is orthogonal, then the output $\mathbf{W}\mathbf{x}_k$ will be proportional to the desired \mathbf{y}_k (see *Linear Algebra for Neural Networks*).

3.2 Delta Rule

An improvement called the *delta rule* can be made in the linear associator; it illustrates a fundamental approach to connectionist learning. The idea is that to define the error, as a function of the weight matrix, as the sum of the differences between the desired and actual outputs of the network, $E(\mathbf{W}) = \sum_{k=1}^p D(\mathbf{y}_k, \mathbf{W}\mathbf{x}_k)$. The weight matrix is then changed by *gradient descent*,

which means that the elements of \mathbf{W} are changed in the relative proportion that causes a maximal incremental decrease of $E(\mathbf{W})$.

If the difference between patterns is measured by Euclidean distance squared, $D(\mathbf{y}, \mathbf{y}') = \|\mathbf{y} - \mathbf{y}'\|^2$ (that is, the sum-of-squares error), then the delta rule is essentially equivalent to linear regression. If the input patterns are linearly independent, then the delta rule will converge to a weight matrix that associates perfectly, $\mathbf{y}_k = \mathbf{W}\mathbf{x}_k$. If they are not, then the weight matrix will be that which minimizes the total error; that is, it will be the best linear prediction of the output patterns from the input patterns (see *Linear Algebra for Neural Networks*).

The delta rule also illustrates an important characteristic of most connectionist networks: their ability to *generalize* to inputs other than those upon which they have been trained. The delta rule provides only linear generalization, but other algorithms, such as backpropagation (Sect. 3.3), can make nonlinear generalizations. Typically, connectionist categories are represented by concrete *prototypes* rather than by definitions in terms of necessary and sufficient conditions or other abstract symbolic structures. Network behavior then depends on similarity to the prototypes rather than on formal manipulation of symbolic structures.

3.3 Backpropagation

Gradient descent may be applied also to multilayer networks of nonlinear units, so long as the activation function is differentiable. The *backpropagation algorithm* (also called the *generalized delta rule*) efficiently computes the weight changes by starting with the last layer and working backward layer by layer. It has been rediscovered a number of times, perhaps first by P. Werbos in 1974, but its importance in connectionism began with its rediscovery in the early 1980s. There are also special adaptations of backpropagation for recurrent networks. Backpropagation has a number of limitations, including: (1) it may be quite slow, (2) it does not necessarily take the shortest path to an error minimum, and (3) it may get trapped in local minima. Nevertheless, it remains a fundamental learning algorithm and has been subject to many practical improvements. *See also*: Backpropagation.

3.4 Other Learning Algorithms

The preceding are examples of *supervised learning* procedures, which means that the “correct answer” \mathbf{y}_k is available for each training input \mathbf{x}_k . Although this is appropriate for modeling some cognitive processes and for many practical problems, for others *unsupervised learning* is preferable. In these procedures, the network is not trained to produce any specific outputs, but it allowed to group or categorize inputs according to standards inherent in its design. Thus unsupervised learning is often equivalent to some kind of statistical clustering. Between these two extremes is *reinforcement learning*, in which the algorithm is told whether or not the output is correct, but not what the correct output is. In

unsupervised and reinforcement learning, as in supervised learning, the network is normally expected to generalize reasonably to novel inputs. There are now hundreds of connectionist learning algorithms, of greater and lesser relevance to cognitive science and neuroscience, but this must suffice for an introduction. *See also:* Artificial Neural Networks: Associative and Self-organizing.

4 Example: Learning Past Tenses

Since the early 1980s, connectionist networks have been used for an enormous number of practical applications and for modeling many aspects of cognition. One notable example must suffice here; see Rumelhart, McClelland & the PDP Research Group (1986) for additional examples.

Rumelhart and McClelland (Rumelhart, McClelland & the PDP Research Group 1986, vol. 2, ch. 18) trained a connectionist network to produce the past tenses of English verbs. The inputs were vectors representing phonological features of the present tenses and the outputs were vectors representing the phonological features of the corresponding past tenses. Learning was observed to pass through three stages. In the first stage, the most common verbs were learned, essentially by rote as individual special cases. In the second stage, the network learned how to form regular past tenses (for it was able to generalize to novel regular verbs), but over-generalized by treating the (previously correctly processed) irregular verbs as though they were regular. In the third stage, the network (re)learned the correct formation of the irregular past tenses without losing its ability to form regular past tenses. It is interesting and highly suggestive that children pass through these same three stages; the model also exhibited errors of the same kinds made by children.

Although this experiment can be (and has been) criticized on a number of grounds as a model of language learning, it is perhaps more valuable as a demonstration of how a connectionist network can exhibit apparently rule-like behavior, but not be following any explicit rules. In particular, exceptions to the rules are handled automatically without explicit accommodation. Thus it is paradigmatic of connectionist cognitive models.

5 Issues

5.1 Connectionist Representations

Connectionist approaches represent and process information in a way that is fundamentally different from *symbolic* approaches, in which knowledge is represented in discrete structures relating atomic lexical-level *features* (i.e., categories of the sort for which natural languages have words). In symbolic approaches, information is processed by formal logic-like rules, which rearrange these atomic units of meaning. In connectionist approaches, on the other hand, information is represented by patterns of activity distributed over large numbers of units, which individually have no lexical-level meaning. The latter are often termed

microfeatures, but they are fundamentally different from *features*, which are supposed to be complete, context-free units of meaning. Microfeatures are just components of distributed representations and are usually individually uninterpretable.

Instead of rules, connectionist information processing is defined by quantitative connections between microfeatures and takes place at a *subsymbolic* level (Smolensky 1988). Cognition then is an emergent effect of large numbers of these interactions (which therefore constitute the *microstructure* of cognition). Indeed, according to this account, apparent symbolic processing is just such an emergent effect of these subsymbolic interactions.

Traditional (symbolic) models of knowledge have been criticized for their *brittleness*. For example, when a set of rules is formulated in an attempt to model the behavior of some expert, it is generally found that the rules do worse than the expert, since the expert applies rules flexibly and makes ad hoc exceptions to them as required by the particulars of the situation. Of course, additional rules can be formulated to cover the exceptions, but then these are likewise found to have exceptions, and so forth. The attempt to reduce flexible behavior to inflexible rules leads to a combinatorial explosion of possibilities which exceeds the capacities of brains and computers.

Connectionist approaches seem better able to account for the flexibility and context-sensitivity of natural intelligence. This is because the connection weights function as a large number of *soft constraints*, none of which are individually necessary or sufficient to produce a result. Therefore connectionist networks can accommodate inputs that are exceptional in various ways, either by ignoring the anomalous aspects, or by corresponding adjustment of their output. As a consequence, their performance is also robust in the face of noise in the input or damage to the network. Further, to the extent that microfeatures of the environment are represented in the input, the network can process information in a way that is sensitive to the context. *See also:* Artificial Intelligence: Connectionist and Symbolic Approaches; Knowledge Representations, Theory of; Production Systems, in Cognitive Psychology; Schemas, Frames and Scripts, in Cognitive Psychology; Symbolic Approaches, in Cognitive Science.

5.2 Criticisms

Connectionist networks have been criticized for their *opacity* or *uninterpretability*. That is, when a connectionist network has been trained to perform some task, it is difficult to extract human-interpretable rules from the network; although the network performs correctly, one cannot understand the “rules” it is apparently following. The reason of course is that it is *not* following rules, and the individual units and weights represent microfeatures and constraints that are lexically meaningless (i.e., have no lexical-level meaning). There are mathematical procedures for extracting rule-like information from networks, but they give only approximations to the network’s behavior. This is analogous to what has been observed during “knowledge acquisition” for expert systems: after the fact, human experts can account in terms of rules for their decision making, but

the rules do not account adequately for the expert's future decisions.

Quite naturally, some of the severest criticisms of connectionism have come from linguists and other cognitive scientists committed to a "language of thought," that is, to the hypothesis that cognition must be understood in terms of the manipulation of propositional or sentential symbolic structures. These criticisms have focused on the alleged inability of "flat" connectionist representations to capture the rich hierarchical symbolic structure of human language and propositional attitudes, and the related sensitivity of cognition to their constituent structure. However, experiments in "connectionist symbol processing" have shown that connectionist networks can be sensitive to the constituent structure of representations without explicit representation of that structure and without the use of explicit symbolic rules (e.g., Sect. 4). There is much more to the issue than this, however, and the early collection by Pinker & Mehler (1988) is still a good introduction to the anti-connectionist position. *See also:* Connectionist Models of Natural Language Processing.

5.3 Computability

If connectionism is viewed as a fundamentally new approach to information representation and processing, then the question arises of its power relative to conventional digital computation, as modeled by the Turing machine. At a basic level this question is easy to answer. On one hand, since connectionist networks are routinely simulated on digital computers, it is apparent that they have no greater power than a Turing machine. On the other hand, researchers have shown that various sorts of connectionist networks can simulate Turing machines, which therefore have no greater power than the networks. The conclusion would seem to be that connectionist networks are equivalent to Turing machines in computing power.

However, at a deeper level the question is problematic, for the Turing machine model is based on certain idealizing assumptions about what is significant and insignificant in models of computation, assumptions that are questionable when applied to connectionist models. In particular, the Turing machine model is based on the assumption that computation proceeds by the recognition of atomic tokens of definite type according to the discrete application of finite, definite rules; these processes are assumed to operate with complete reliability. These assumptions are a poor match to connectionist approaches, in which information is represented in distributed patterns of continuous activity, and in which recognition is a matter of degree. Nevertheless, questions of computability are also important in connectionism, but relevant answers may require the development of a new theory of computation that makes idealizing assumptions more relevant to connectionist approaches. *See also:* Theory of Computation.

5.4 Relation to Biological Neural Networks

Connectionist networks are often called "neural networks" and described in terms of (artificial) neurons connected by (artificial) synapses, but is this more

than a metaphor? Generally, connectionist models have reflected the contemporary understanding of neurons. For example, McCulloch and Pitts focused on the “all or nothing” character of neuron firing, and modeled neurons as digital logic gates. Newer connectionist models have had a more analog focus, and so the activity level of a unit is often identified with the instantaneous firing rate of a neuron. However, these models still ignore many important properties of real neurons, which may be relevant to neural information processing (Rumelhart, McClelland & the PDP Research Group 1986, vol. 2, ch. 20). As a consequence neuroscientists have stressed the differences between biological neurons and the simple units in connectionist networks; the relation between the two remains an open problem. Nevertheless, it is much easier to envision neural implementations of connectionist networks than of symbol-processing architectures.

See Churchland (1986) and Quinlan (1991) for an introduction to connectionist approaches in philosophy and psychology. *See also*: Connectionist Models of Concept Learning; Connectionist Models of Development.

References

- Anderson, James A., Andras Pellionisz & Edward Rosenfeld, eds. 1990. *Neurocomputing 2: Directions for Research*. Cambridge MA: MIT Press.
- Anderson, James E. & Edward Rosenfeld, eds. 1988. *Neurocomputing: Foundations of Research*. Cambridge MA: MIT Press.
- Churchland, Patricia Smith. 1986. *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge MA: MIT Press.
- Hinton, Geoffrey E. & James A. Anderson, eds. 1989. *Parallel Models of Associative Memory*. updated ed. Hillsdale NJ: Lawrence Erlbaum Associates.
- Pinker, Steven & Jacques Mehler, eds. 1988. *Connections and Symbols*. Cambridge MA: MIT Press.
- Quinlan, Philip. 1991. *Connectionism and Psychology: A Psychological Perspective on New Connectionist Research*. Chicago IL: University of Chicago Press.
- Rumelhart, David E., James L. McClelland & the PDP Research Group. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge MA: MIT Press.
- Smolensky, P. 1988. “On the Proper Treatment of Connectionism.” *Behavioral and Brain Sciences* 11:1–74.