Connectionist Modeling

David C. Plaut
Departments of Psychology and Computer Science, Carnegie Mellon University,
and the Center for the Neural Basis of Cognition

January 1999

To appear in Kasdin, A. (Ed.), *Encyclopedia of Psychology*. Washington DC:
Americal Psychological Association.

Since the pioneering work of Alan Newell and Herb Simon in the late 1950s, researchers interested in human cognitive processes have used computer simulations to try to identify general principles of cognition. The strategy is to build computational models that embody putative principles and then examine how well the models capture human performance in cognitive tasks. Until the 1980s, this effort was undertaken largely within the context of the "computer metaphor" of mind: Researchers thought of the human mind as though it were a conventional digital computer and built computational models based on this conceptualization. Indeed, this approach has lead to considerable progress in modeling explicit reasoning, problem solving, and other high-level cognitive processes.

By the late 70s and early 80s, however, many researchers began to think that an alternative framework was needed to capture the full range of cognitive behavior—one based more closely on the style of computation employed by the brain. The new approach, called connectionist or neural network modeling, or the parallel distributed processing approach (Elman et al., 1996; McLeod et al., 1998; Rumelhart et al., 1986), implements cognitive processes in terms of massively parallel cooperative and competitive interactions among large numbers of simple neuron-like computational units. Unit interactions are governed by modifiable excitatory and inhibitory weights on connections among the units. Although each unit exhibits nonlinear spatial and temporal summation, units and connections are not generally be taken as corresponding directly to individual neurons and synapses. Rather, the connectionist approach attempts capture the essential computational properties of the vast ensembles of real neuronal elements found in the brain using simulations of smaller networks of more abstract units. By linking neural computation to behavior, the framework enables developmental, cognitive and neurobiological issues to be addressed within a single, integrated formalism.

REPRESENTATION. An issue of central relevance in understanding cognition is the nature of the representations used in cognitive processes. There are two basic approaches to representation within connectionist networks.

1. In a localist representation, familiar entities such as letters, words, concepts, and propositions are encoded by the activity of individual units.

2. In a distributed representation, such entities are encoded by alternative patterns of activity over the same units, such that each entity is represented by the activity of many units and each unit participates in representing many entities.

Models employing localist representations are sometimes termed structured networks, although this should not be taken to imply that models using distributed representations are unstructured.

Many early influential connectionist models in psychology employed localist representations. For example, the Interactive Activation model (McClelland & Rumelhart, 1981) consisted of three layers of units: letter-feature units, letter units, and word units. Units in each layer received excitatory connections from consistent units at other layers and inhibitory connections from inconsistent alternatives within the same layer. The resulting interactive processing played a critical role in explaining a number of context effects in perception, including the word superiority effect, in which the perception of a letter is enhanced

when it occurs in the context of a word compared with when it occurs in isolation or in a random letter string. An analogous model in the domain of spoken word recognition, called TRACE (McClelland & Elman, Chapter 15 of Rumelhart et al., 1986), accounts for similar phenomena.

Most recent connectionist models rely on properties of distributed representations. Although distributed representations can be less intuitive, they are attractive in part because they provide a more natural account of the richness and subtlety of the relationships among entities. The key to distributed representations is the use of patterns whose similarity relations capture similarities in the roles the patterns play in cognition, since, in connectionist models, similar patterns have similar consequences (Hinton, McClelland & Rumelhart, Chapter 3 of Rumelhart et al., 1986). Distributed representations can also be used to implement more complex, relational knowledge structures like frames and scripts if units encode conjunctions of roles and properties of role-fillers—in fact, such representations emerge naturally when networks are trained on tasks in which entities enter into multiple types of relations (see Hinton, 1991).

LEARNING. Most distributed connectionist models place strong emphasis on learning, in part because it is difficult to hand-specify effective sets of weights in such systems. Learning in connectionist networks involves modifying the weights on connections in a way that influences the pattern of unit activations produced in response to a given input. There are three broad frameworks for learning.

1. Supervised learning involves changing weights so as to reduce the discrepancy between the actual output generated by the system for a given input and the correct output, which is assumed to be provided by an external "teacher."

2. Unsupervised learning involves changing weights based only on the input provided to the system and intrinsic biases built into the learning procedure, without any explicit feedback based on the behavior of the system.

3. Reinforcement learning is something of a middle ground; it involves changing weights based on minimal performance feedback—typically only a scalar value indicating the "goodness" of outcomes that depend on the behavior of the system.

Whereas unsupervised and reinforcement learning are more directly related to known learning mechanisms in the brain, the majority of applications of connectionist modeling in cognitive psychology have employed supervised learning. This is because supervised learning is more effective at developing internal representations that can support the complex transformations involved in many forms of cognitive processing.

The most commonly employed form of supervised learning is back-propagation (Rumelhart, Hinton & Williams, Chapter 8 of Rumelhart et al., 1986). This procedure involves iteratively 1) computing activations in a forward pass from input units to output units, possibly via one or more layers of hidden units (so called because they are not visible to the environment); 2) computing a measure of performance error over the output units, 3) propagating this error backward through the network (using the chain rule from calculus) to determine the partial derivative of the error with respect to each weight in the network; and finally 4) changing the weights based on these derivatives so as to reduce the error. Although it is highly unlikely that the brain employs back-propagation in its literal form, there are more biologically plausible procedures (see, e.g., O'Reilly, 1996) which are computationally equivalent (albeit somewhat less efficient).

In an early application of error-correcting learning, Rumelhart and McClelland (1986) showed that a single network could learn to generate the past tense forms of both regular and irregular English verbs from their stems, thereby obviating the need for dual rule-based and exception mechanisms. Although aspects of the approach were strongly criticized (Pinker & Prince, 1988), many of the specific limitations of the model have been addressed in subsequent simulation work. A similar line of progress has taken

place in the domain of English word reading (see Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989).

PROCESSING. The architecture of a network refers to the numbers of units it contains, how they are organized into groups or layers, and how these layers are interconnected. There are three general classes of architecture used widely in connectionist modeling.

1. A feedforward network consists of a series of layers of units with a restricted pattern of connectivity, such that units project only to later layers, never within a layer or back to earlier layers.

2. A fully recurrent network has no restriction on connectivity, so that any unit may potentially be connected to any other unit (including itself).

3. A simple recurrent network is something of a hybrid: Processing is feedforward over a series of layers but the states of certain "context" input layers are set by copying the previous states of some hidden or output units. These context states allow the network to learn to be sensitive to temporal dependencies among successive inputs (Elman, 1990).

In practice, many distributed models employing learning have a feedforward or simple recurrent architecture, whereas non-learning models like the Interactive Activation model are typically fully recurrent. This is, however, more a computational convenience than a theoretical discrepancy, as the versions of back-propagation that are applicable to fully recurrent networks require far greater computational resources.

Even so, fully recurrent networks are increasingly being applied directly in modeling psychological phenomena. Most of these models are attractor networks, in which units interact to cause the network as a whole to settle gradually into a stable pattern of activity corresponding to the network's interpretation of the input. Attractor networks are particularly appropriate for modeling processes that involve selecting among alternatives, such as word recognition and comprehension (Hinton & Shallice, 1991).

Although fully recurrent networks are capable of learning to exhibit more complex temporal behavior, for reasons of efficiency it is more common to apply simple recurrent networks in temporal domains. For example, Elman (1991) demonstrated that a simple recurrent network could learn the structure of an English-like grammar, involving number agreement and variable verb argument structure across multiple levels of embedding, by repeatedly attempting to predict the next word in processing sentences. St. John and McClelland (in Hinton, 1991) also showed, for a somewhat simpler corpus, how such networks can learn to develop a representation of sentence meaning by attempting to answer queries thematic role assignments throughout the course of processing a sentence.

CONCLUSION. The connectionist framework for modeling human cognition has led to the development of explicit computational models of a wide range of cognitive functions. In many cases, these models introduce new ways of thinking about the nature of the computations that are performed and and how learning can give rise to the ability to carry out these computations. The models also give us new ways of relating cognitive processes to brain function. Connectionist models will play an increasingly important role in the development of cognitive theories that are both mechanistically explicit and neurobiologically realistic.

Annotated Bibliography

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7,* 195–225.

These two articles by Elman introduce simple recurrent networks and provide a number of demonstations of how they can be applied in learning complex temporal tasks.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

Articulates and illustrates how connectionist modeling can provide important insights into cognitive development without invoking domain-specific innate constraints.

Hinton, G. E. (Ed.). (1991). *Connectionist symbol processing.* Cambridge, MA: MIT Press.

A collection of important papers on applying connectionist networks to issues in higher-level cognition.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review, 98,* 74–95.

The first major attempt to use connectionist networks to explain cognitive impairments due to brain damage.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88,* 375–407.

A major landmark in the application of computational modeling in cognitive psychology. Many current-day models of letter and word perception retain much of the basic structure of this model.

McLeod, P., Plunkett, K., & Rolls, E. T. (1998). *Introduction to connectionist modelling of cognitive processes.* Oxford, UK: Oxford University Press.

A recent introductory text on connectionist modeling in psychology. Includes an easy-to-use software package and extensive examples for hands-on experience. A good follow-up to the Rumelhart, McClelland and the PDP Reserach Group two-volume set on Parallel Distributed Processing.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation, 8,* 895–938.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28,* 73–193.

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

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (pp. 216–271). Cambridge, MA: MIT Press.

Presents one of the most influential and controversial connectionist models—applied to the generation of the past tense of English verbs.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations.* McClelland,

J. L., Rumelhart, D. E., and the PDP Research Group (Eds.), *Volume 2: Psychological and biological models.* Cambridge, MA: MIT Press.

This two-volume set provided the first systematic exploration of how connectionist models could be applied effectively across a wide range of cognitive domains, as well as the theoretical and mathematical background necessary to understand the approach. Although a bit dated, it is still the best place to start in gaining an understanding of connectionist cognitive modeling.

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