Connectionism: Past, Present, and Future

Jordan B. Pollack
Computer & Information Science Department
The Ohio State University

1. Introduction

Research efforts to study computation and cognitive modeling on neurally-inspired mechanisms have come to be called Connectionism. Rather than being brand-new, it is actually the rebirth of a research program which thrived from the 40's through the 60's and then was severely retrenched in the 70's. Connectionism is often posed as a paradigmatic competitor to the Symbolic Processing tradition of Artificial Intelligence (Dreyfus & Dreyfus, 1988), and, indeed, the counterpoint in the timing of their intellectual and commercial fortunes may lead one to believe that research in cognition is merely a zero-sum game. This paper surveys the history of the field, often in relation to AI, discusses its current successes and failures, and makes some predictions for where it might lead in the future.

2. Early Endeavors: High Hopes and Hubris

Before the explosion of symbolic artificial intelligence, there were many researchers working on mathematical models of intelligence inspired by what was known about the architecture of the brain. Under an assumption that the mind arises out of the brain, a reasonable research path to the artificial mind was by simulating the brain to see what kind of mind could be created. At the basis of this program was the assumption that neurons were the information processing primitives of the brain, and that reasonable models of neurons connected into networks would suffice.

2.1. McCullogh & Pitts

The opening shot in neural network research was the 1943 paper by Warren S. McCullogh and Walter Pitts. In ''A logical calculus of ideas immanent in nervous activity'' they proved that any logical expression could be ''implemented'' by an appropriate net of simplified neurons.

They assumed that each ''neuron'' was binary and had a finite threshold, that each ''synapse'' was either excitatory or inhibitory and caused a finite delay (of 1 cycle), and that networks could be constructed with multiple synapses between any pair of nodes. In order to show that any logical expression is computable, all that is necessary is to build the functions AND, OR and NOT. Figure 1 shows the simple networks which accomplish these functions. And in order to build ''larger'' functions, one need only glue these primitive together. For example, figure 2 shows a two layer network which computes the exclusive-or function. Continuing in this vein, one could construct a computer by building up the functional parts, e.g. memories, ALU's, from smaller pieces, which is exactly how computers are built.

McCullogh and Pitts proved several theorems about equivalences of different processing assumptions, both for simple nets and for nets with feedback cycles, using a somewhat arcane syntax of temporal propositions. Since learning was not under consideration, memory, for them, was based on ''activity [which] may be set up in a circuit and continue reverberating around it for an indefinite period of time''[1]. They concluded with a discussion of Turing computability, about which we will have more to say in chapter 4.

2.2. Hebb

There was very little psychology in the science of neural nets, and very few neural considerations in the mainly Stimulus-Response psychology of the day. In *The Organization of Behavior*, Donald O. Hebb set out to rectify this situation, by developing a physiologically-motivated theory of psychology.

Rejecting reflexes, Hebb put forth and defended the notion of an *autonomous central process*, which intervenes between sensory input and motor output. Of his own work he said:

> The theory is evidently a form of *connectionism*, one of the switchboard variety, though it does not deal in direct connections between afferent and efferent pathways: not an ''S-R'' psychology if R means a muscular response. The connections serve rather to establish autonomous central activities, which then are the basis of further learning.[2]

---

[1] (McCullogh & Pitts, 1943), p. 124.

[2] Hebb (1949, p. xix), emphasis mine.

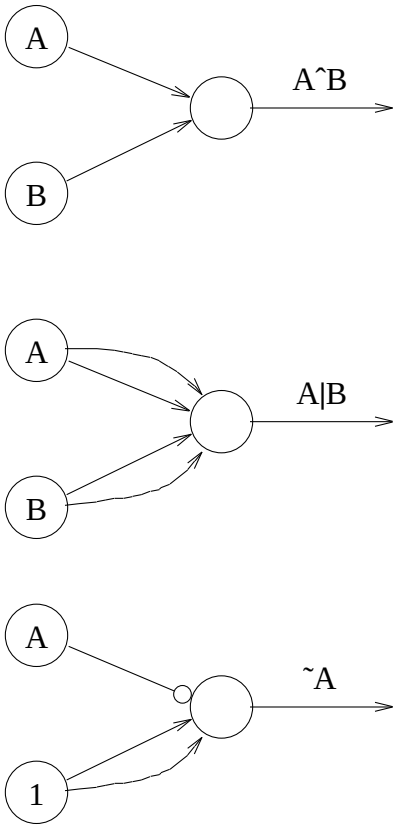
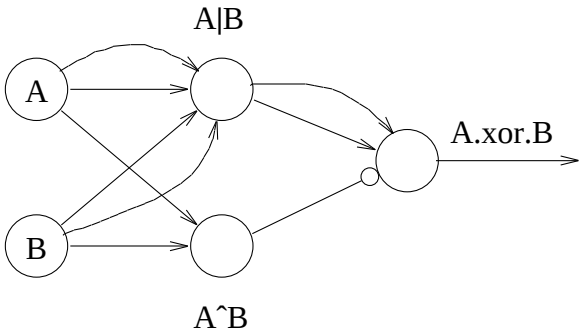


Figure 1.

Logical primitives AND, OR, and NOT implemented with McCullogh & Pitts neurons. A neuron ‘‘fires’’ if it has at least two activating synapses (arrow links) and no inhibiting inputs (circle links).

Figure 2.



A two-layer McCullogh-Pitts network which computes exclusive-or as the function A^B-A|B.

Of an incredibly rich work, Hebb is generally credited with two notions that continue to hold influence on research today. The first is that memory is stored in connections and that learning takes place by synaptic modification:

> Let us assume then that the persistence or repetition of a reverberatory activity (or ‘‘trace’’) tends to induce lasting cellular changes that add to its stability. The assumption can be precisely stated as follows: *When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A’s efficiency, as one of the cells firing B, is increased.*[3]

[3] Ibid., p. 62.

And the second is that neurons don't work alone, but may, through learning, become organized into larger configurations, or ''cell-assemblies'', which could thus perform more complex information processing.

2.3. Ashby

In *Design for a Brain*, W. Ross Ashby laid out a methodology for studying adaptive systems, a class of machines to which, he asserted, the brain belongs. He set out an ambitious program:

> [...] we must suppose (and the author accepts) that a real solution of our problem will enable an artificial system to be made that will be able, like the living brain, to develop adaptation in its behaviour. Thus the work, if successful, will contain (at least by implication) a specification for building an artificial brain that will be similarly self-co-ordinating.[4]

While the work was not ''successful'' in these terms, Ashby laid the groundwork for research that is flourishing today. His methodology for studying dynamical systems as fields of variables over time is echoed today in the connectionist studies which involve time evolution of dynamical systems (Hopfield, 1982; Smolensky, 1986) and his notion of building intelligent machines out of homeostatic elements can be seen as precursor to Klopf's (1982) work on heterostatic elements.

2.4. Rosenblatt

Hebb's notion of synaptic modification was not specified completely enough to be simulated or analyzed. Frank Rosenblatt studied a simple neurally-inspired model, called a *perceptron*, for many years, and summarized his work in a 1962 epic, *Principles of Neurodynamics*.

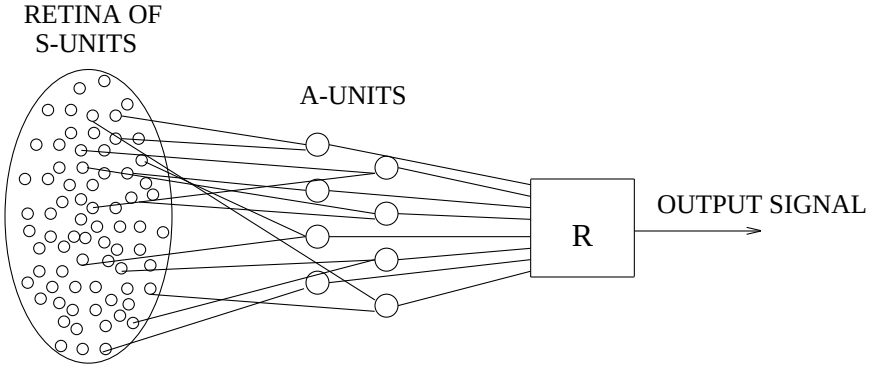


Figure 3.

An elementary perceptron, which consists of a feedforward network from a set (retina) of input units (S-Units) connected with fixed weights to a set of threshold units (A-Units) connected with variable weights to an output unit (R-Unit).

Rather than using the fixed weights and thresholds and absolute inhibition of the McCullogh-Pitts neuron, Rosenblatt's units used variable weights with relative inhibition. A perceptron consisted of many such units arranged into a network with some fixed and some variable weights. Figure 3 shows a typical elementary perceptron. Usually used for pattern-recognition tasks such as object classification, an elementary perceptron consisted of a ''retina'' of binary inputs, a set of specific feature detectors, and a response unit. The weights from the input to the middle layer were fixed for an application, and the weights from the detectors to the response unit were iteratively adjusted. The major results of Rosenblatt's work were procedures for adjusting these variable weights on various perceptron implementations, conditions of existence for classification solutions, and proofs that these procedures, under the right conditions, converged in finite time. One statement of the famous ''perceptron convergence theorem'' from Rosenblatt is as follows:

> Given an elementary α -perceptron, a stimulus world W, and any classification $C(W)$ for which a solution exists; let all stimuli in W occur in any sequence, provided that each stimulus must reoccur in finite time; then beginning from an arbitrary initial state, an error correction procedure (quantized or non-quantized) will always yield a solution to $C(W)$ in finite time, with all signals to the R-unit having magnitude at least equal to an arbitrary quantity $\delta \geq 0$.[5]

A world consisted of a set of input patterns to the retina, and a classification was a separation of this world into positive and negative classes. The existence of a guaranteed convergence procedure was very

---

[4] Ashby (1960, p. 10).

useful; the rub was that that the kinds of classifications ‘‘for which a solution exists’’ were extremely limited. As a footnote to the somewhat incomprehensible proof of this theorem, Rosenblatt attacked a shorter alternative proof by Seymour Papert; an attack, we are sure, he eventually regretted.

3. Symbolic Seventies: Paying the Price

Many of the early workers of the field were given to extravagant or exuberant claims or overly ambitious goals. McCullogh & Pitts, for example, asserted that ‘‘specification of the net would contribute all that could be achieved in [psychology]’’, Ashby clearly overestimated the power of his homeostat, and Rosenblatt stated at a 1958 symposium that:

> [...] it seems clear that the Class C’ perceptron introduces a new kind of information processing automaton: for the first time we have a machine which is capable of having original ideas.[6]

He made other dubious claims of power for his perceptrons as well, which undoubtedly provoked a backlash. Discussions of this controversy can be found in (Rumelhart & Zipser, 1986) or (Dreyfus & Dreyfus, 1988), and some interesting perspectives on some of the personalities involved can be found in chapter 4 of McCorduck (1979).

3.1. Minsky & Papert

> *I was trying to concentrate on a certain problem but was getting bored and sleepy. Then I imagined that one of my competitors, Professor Challenger, was about to solve the same problem. An angry wish to frustrate Challenger then kept me working on the problem for a while...*[7]

In 1969, Marvin Minsky and Seymour Papert published *Perceptrons*, a tract which sounded the deathbell for research on perceptrons and other related models. A thoroughgoing mathematical analysis of linear threshold functions showed the limitations of perceptrons both as pattern-recognition machines and as general computational devices.

This book will probably stand permanently as one of the most important works in the field of connectionism, so it is important to understand some of the findings of Minsky & Papert.

First, they defined the *order* of a predicate as the size of the largest conjunction in the minimal sum-of-products logical form for that predicate (or its inverse). Thus while both conjunction and alternation are predicates of order 1, exclusive-or is a predicate of order 2. The generalization of exclusive-or to more than 2 inputs is parity, which is not of finite order: a sum of products to represent parity of n inputs has at least one term of size n. As a predicate of non-finite order is scaled, then, there is no limit to the necessary fan-in of units, and perceptrons lose their nice aspect of locality.

Using arguments of symmetry, Minsky & Papert then showed, with their *Group Invariance Theorem*, that linear threshold functions which are invariant under a permutation group can be transformed into a function whose coefficients depend only on the group; a major result is that the only linear (i.e. order 1) functions invariant under such transitive groups as scaling, translation and rotation are simple size or area measures. Attempts to use linear functions for, say, optical character recognition under these transitive conditions are thus doomed to failure.

After a cataloging of the orders of various geometric functions, Minsky & Papert focused on the problems of learning. They showed that as various predicates scale, the sizes of coefficients can grow exponentially, thus leading to systems of impractical memory requirements needing unbounded cycles of a convergence procedure. As Rumelhart & Zipzer pointed out in their review of the perceptron controversy:

> The central theme of [*Perceptrons*] is that parallel recognizing elements, such as perceptrons, are beset by the same problems of scale as serial pattern recognizers. Combinatorial explosion catches you sooner or later, although sometimes in different ways in parallel than in serial.[8]

Minsky & Papert worked on the problem of perceptrons for quite a long time, the result being a boring and sleepy decade for neurally-inspired modeling.

3.2. Everybody Else

Despite the herbicidal effect of *Perceptrons* on neural network research funding and the flowering of symbolic AI, some research efforts continued to grow during the 70’s. Neural network researchers just could not easily publish their work in the AI journals or conferences.

---

[5] Rosenblatt (1962, p. 111).

[6] Rosenblatt (1959, p. 449). This type of claim has not been repeated in AI history until the advent of ‘‘Discovery systems’’ (Lenat, 1977).

[7] Minsky (1986, p. 42).

[8] Rumelhart & Zipser (1986, p. 158).

A lot of the work dealt with associative or content addressable memories. Though beyond the scope of this history, significant developments and analyses can be found in the works of Teuvo Kohonen (Kohonen, 1977; Kohonen et al., 1981) and David Willshaw (Willshaw, 1981).

(Anderson et al., 1977) described experiments with a saturating linear model for pattern association and learning called the ''Brain-State in a Box'' or BSB model. Given the current state of the system as a vector, $\vec{X}(t)$ and a matrix of weights, W, the next state of the system can be computed as the inner product between the state and weights, bounded between -1 and 1:

$$\vec{X}(t+1) = \min(1, \max(-1, \vec{X}(t) + W \cdot \vec{X}(t)))$$

Under this system, the state of the system is always within an n-dimensional hypercube (i.e. a ''box'') centered around the origin. Anderson was able to apply a type of Hebbian associative learning rule to find weights for this system. BSB models are still being used productively, for example, in the lexical access model of (Kawamoto, 1985).

It is almost impossible to quantify the huge contribution of Stephen Grossberg to neural modelling. The scholarly output of Grossberg and his colleagues at Boston University's Center for Adaptive Systems throughout the seventies is daunting in its mathematical sophistication. Though no excuse, this might account for the allegedly poor scholarship on the part of modern connectionists:

> [Rumelhart & Zipser's] discussion does not, however, acknowledge that both the levels and the interactions of a competitive learning model are incompatible with those of an interactive activation model (Grossberg 1984). The authors likewise do not state that the particular competitive learning model which they have primarily analyzed is identical to the model introduced and analysed in Grossberg (1976a, 1976b), nor that this model was consistently embedded into an adaptive resonance model in Grossberg (1976c) and later developed in Grossberg (1978) to articulate the key functional properties [of interactive activation] which McClelland & Rumelhart (1981) describe...[9]

It is, of course, possible that the Connectionism of the 80's might in the future be seen as ''Grossberg: Rediscovered''.

4. Exhuberant Eighties: Research Reborn

Interest in connectionist modeling has been on the rise in the 1980's. Perhaps the limits of the symbolic paradigm were beginning to show, perhaps the question of how to program parallel computers became more relevant as their construction became cost-effective, perhaps some agency simply began funding neural models, or perhaps it was simply the ebb and flow of scientific interest. Whatever the reason, the rebirth is now in full swing. This section reviews some of the highlights of recent connectionist history.

4.1. Interactive Activation

UCSD's Center for Human Information Processing, one of the nation's leading centers for cognitive science, was a staunch supporter of the symbolic paradigm in information-processing psychology. With the publication of *Explorations in Cognition* in 1974, David Rumelhart and Don Norman laid out a research program strictly in line with the main elements of AI of the time. Propositions, Procedures, Semantic Networks, and Augmented Transition Networks were all used in service of a theory of psychology, and actual computer programs were built which supported the theory.

In 1980, a pair of curious reports were issued from the center: ''An interactive activation model of the effects of context in perception, parts 1 and 2'' by David Rumelhart and James McClelland (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982). Gone were the labelled semantic networks, propositions and procedures. Gone was the link to mainstream AI. Instead, there were ''neuron-like'' units, communicating through spreading activation and lateral inhibition. Basically a very small model for explaining many well-known psychological effects of letter recognition in the context of words, their interactive activation model was one of the first high-profile successful applications of modern connectionism.

McClelland & Rumelhart's system, which simulated reactions to visual displays of words and nonwords, dealt only with 4-letter words, and was organized into three distinct levels, **word, letter,** and **feature**. The word level contained a group of 1179 units, one for each word, the letter level contained 4 groups of 26 units each, and the feature level contained 4 groups of 12 units each for stylized visual features of letters. The system operated by providing input to visual features in all 4 letter positions; this activation caused activation in various letter units, which, in turn, caused the activation of possible words. Each group of letter units, and the group of word units, formed what are now called ''winner-take-all''

---

[9] Grossberg (1987, pp. 27-28).

networks, by being fully connected with lateral inhibition links, so that a single unit would tend to dominate all others. Finally, the word units gave positive feedback to their corresponding 4 letter units.

Clearly in the class of programmed, as opposed to trained, neural network models, Rumelhart & McClelland avoided the morass of individually assigning weights by using uniform weighting schemes for each class of links.

They also provided a justification for the constraints on their model, a justification which neatly sidesteps any claim of neural reality that could open up a philosophical can of worms:

> We have adopted the approach of formulating the model in terms which are similar to the way in which such a process might actually be carried out in a neural or neural-like system. We do not mean to imply that the nodes in our system are necessarily related to the behavior of individual neurons. We will, however, argue that we have kept the kinds of processing involved well within the bounds of capability for simple neural circuits.[10]

4.2. The Clarion Call

In 1982, Jerry Feldman & Dana Ballard published ''Connectionist Models and their Properties'', a focusing paper which helped to legitimize connectionism as a methodology for AI and cognitive science. Drawing on both their own work in vision and related neurally-inspired models such as the Rumelhart & McClelland work mentioned above, they sought to unify several strands of research in different fields and define (and name) the bandwagon.[11] Their justifications for abandoning symbolic AI and taking up connectionism were fourfold. First, animal brains are organized differently than computers. Second,

> Neurons whose basic computational speed is a few milliseconds must be made to account for complex behaviors which are carried out in a few hundred milliseconds. This means that *entire complex behaviors are carried out in less than a hundred time steps.*[12]

Third, by studying connectionism we may learn ways of programming the massively parallel machines of the future. And, fourth, many possible mechanisms underlying intelligent behavior cannot be studied within the symbolic programming paradigm.

Feldman & Ballard painted the possibilities of parallelism with broad brushstrokes. Using a framework which included both digital and analog computation, they offered up a large bag of tricks including both primitives for constructing systems (Winner-Take-All Networks, and Conjunctive Connections) and organizing principles to avoid the inevitable combinatorial explosion (Functional Decomposition, Limited Precision Computation, Coding, and Tuning). Although their paper was sprinkled with somewhat fanciful examples, the successful application of their ''tricks'' can be seen in several of the dissertations produced by their students (Cottrell, 1985b; Sabbah, 1982; Shastri, 1988).

4.3. Hopfield Nets

One of the interesting sociological aspects of the rebirth of connectionism is that valuable contributions are being made from areas other than computer science and psychology. There are several ideas from physics which have entered into the discussion, and perhaps the most notable contributions have come from J. J. Hopfield.

In (Hopfield, 1982), he laid out a system for building associative memories based on an analogy to a well-studied physical system, spin glasses. Hopfield showed that, by using an asynchronous and stochastic method of updating binary activation values, local minima (as opposed to oscillations) would reliably be found by his method:

> Any physical system whose dynamics in phase space is dominated by a substantial number of locally stable states to which it is attracted can therefore be regarded as a general content-addressable memory. The physical system will be a potentially useful memory if, in addition, any prescribed set of states can readily be made the stable states of the system.[13]

Hopfield devised a novel way of ''bulk programming'' a neural model of associative memory by viewing each memory as a local minimum for a global ''energy'' function. A simple computatation converted a set of memory vectors into a symmetric weight matrix for his networks.

---

[10] McClelland & Rumelhart (1981, p. 387).

[11] ''Connectionism'', was the name of some very antiquated psychological theory that even Hebb alluded to. And though the Rumelhart School has tried to rename their work as ''Parallel Distributed Processing'' or ''PDP'' models, it seems that ''Connectionism'' (or ''Connexionism'' in Britain) has stuck. But new names keep appearing, such as Neurocomputing, Neuro-engineering, and Artificial Neural Systems...

[12] Feldman & Ballard (1982, p. 206.)

[13] Hopfield (1982, p. 2554).

In (Hopfield & Tank, 1985) he extended his technique of bulk programming of weights to analog devices and applied it to the solution of optimization problems, such as the NP-complete Travelling Salesman problem. By designing an energy function whose local minima (or ''attractor states'') corresponded to good circuits for a particular configuration of cities, Hopfield's network could rapidly find a reasonably good solution from a random initial state. It should be noted that Hopfield's motivation was **not** to suggest the possibility that $P = NP$ nor to introduce a new approximate algorithm for NP-complete problems, but to demonstrate the usefulness of his neural networks for the kinds of problems which may arise for ''biological computation'', an understanding of which may ''lead to solutions for related problems in robotics and data processing using non-biological hardware and software.''[14]

Hopfield has become the symbolic founding father of a very large and broad physics-based study of neural networks as dynamical systems, which is beyond the scope of this survey.

4.4. Born-Again Perceptrons

Of the extension of perceptron learning procedures to more powerful, multilayered systems, Minsky & Papert said:

> We consider it to be an important research problem to elucidate (or reject) our intuitive judgement that the extension is sterile. Perhaps some powerful convergence theorem will be discovered, or some profound reason for the failure to produce an interesting ''learning theorem'' for the multilayered machine will be found.[15]

In the past few years, however, several techniques have appeared which seem to hold the promise for learning in multilevel systems. These are (1) Associative Reward-Penalty, (2) Boltzmann Machine Learning, and (3) Back Propagation.

4.4.1. Associative Reward-Penalty

Working with the goal-seeking units of (Klopf, 1982), Andrew Barto and colleagues published results in 1982 on one of the first perceptron-like networks to break the linear learning barrier (Barto et al., 1982). Using a two-layered feed-forward network they demonstrated a system which learned to navigate towards either of 2 locational goals in a small landscape. They showed that in order to have done this successfully, the system had to essentially learn exclusive-or, a nonlinear function.

The task was posed as a control problem for a ''simple organism'': At any time t the input to the network was a 7-element vector indirectly indicating location on a two dimensional surface. The output of the network was 4 bits indicating which direction to move (i.e. north, east, south or west). And a reinforcement signal, broadcast to all units, based on the before/after difference in distance from the goals, was used to correct the weights.

The network had 8 ''hidden'' units interposed between the 7 input units and 4 output units. One of the factors contributing to the success of their method was that instead of the hidden layer computing binary thresholds, as in an elementary perceptron, it computed positive real numbers, thus allowing gentler gradients for learning.

This early work, on a specific network with a few quirks, was subsequently developed into into a more general model of learning, the *Associative Reward-Penalty* or A_{R-P} algorithm. See (Barto, 1985) for an overview of the work.

4.4.2. Boltzmann Machines

> **Anneal** - To toughen anything, made brittle from the action of fire, by exposure to continuous and slowly diminished heat, or by other equivalent process.
>
> *You have been wasted one moment by the vertical rays of the sun and the next annealed hissing hot by the salt sea spray.*[16]

Another notion from physics which has been ported into connectionism is *simulated annealing*. Based on the work of (Kirkpatrick et al., 1983), Ackley, Hinton, & Sejnowski (1985) devised an iterative connectionist network which relaxes into a global minimum. As mentioned earlier, Hopfield (1982) constructed a network for associative memory (in which each memory was a local minimum) and showed that an asynchronous update procedure was guaranteed to find local minima. By utilizing a simulated annealing procedure, on the other hand, a Boltzmann Machine[17] can find a global minimum.

---

[14] Hopfield & Tank (1985, p. 142).

[15] Minsky & Papert (1969, p. 232).

[16] Definition from the compact edition of the Oxford English Dictionary; their citation from Michael Scott, *Cringle's Log*, 1859.

[17] No hardware actually exists. The name is an honor, like Von Neumann Machine.

Given a set of units, s_i, which take on binary values, connected with symmetric weights, w_{ij} , the overall ''energy'' of a particular configuration is:

$$E = -\sum_{i<j} w_{ij} s_i s_j + \sum_i \theta_i s_i$$

where θ_i is a threshold. A local decision can be made as to whether or not a unit should be on or off to minimize this energy. If a unit is off (0), it contributes nothing to the above equation, but it it is on (1), it contributes:

$$\Delta E_i = \sum_j w_{ij} s_j - \theta_i$$

In order to minimize the overall energy, then, a unit should turn on if its input exceeds its threshold and off otherwise.

But because of the interaction of all the units, a simple deterministic or greedy algorithm will not work. The Boltzmann machine uses a stochastic method, where the probability of a unit's next state being on is:

$$\frac{1}{1+e^{-\Delta E_k/T}}$$

where T is a global ''temperature'' constant. As this temperature is lowered toward 0, the system state freezes into a particular configuration:

> At high temperatures, the network will ignore small energy differences and will rapidly approach equilibrium. In doing so, it will perform a search of the coarse overall structure of the space of global states, and will find a good minimum at that coarse level. As the temperature is lowered, it will begin to respond to smaller energy differences and will find one of the better minima within the coarse-scale minimum it discovered at high temperature.[18]

To use simulated annealing as an iterative activation function, some units must be ''clamped'' to particular states, and a ''schedule'' of temperatures and times is used to drive the system to ''equilibrium''. This type of relaxation has been used in two parsing models so far, (Selman, 1985) and (Sampson, 1986), and is a computational primitive in the connectionist production system of (Touretzky & Hinton, 1985).

The real beauty of the Boltzmann Machine comes through in its very simple learning rule. Given a desired set of partial states to learn and an initial set of weights, the learning procedure, using only local information, can interactively adjust the weights. By running the annealing procedure several times while clamping over the learning set and several times without any clamping, statistical information about how to change all the weights in the system can be gathered. With slow annealing schedules, their procedure can learn codings for hidden units, thus overcoming some of the limitations of perceptrons.

The down-side of all this is that the learning algorithm is very slow and computationally expensive. Learning a set of weights for a problem may take only hundreds of iterations - but each iteration, in order to collect the statistical information, consists of several trials of simulated annealing, possibly with gentle schedules of thousands of temperatures, for each test case.

4.4.3. Back-Propagation

A more robust procedure for learning in multiple levels of perceptron-like units was independently invented and reinvented by several people. In 1981, David Parker apparently disclosed self-organizing logic gates to Stanford University with an eye towards patenting, and he reported his invention in (Parker, 1985); Parker also recently discovered that Paul Werbos developed it in a 1974 mathematics thesis from Harvard University. Yann Le Cun (1985) described a similar procedure in French, and Rumelhart, Hinton, & Williams (1986) reported their method, finally, in English.

Perceptrons were linear threshold units in two layers: The first layer detects a set of features, which were hard-coded; the second layer linearly combined these features and could be trained. Convergence procedures for perceptrons would only work on one layer, however, which, among other problems, severely limited their usefulness.

One explanation for why learning could not be extended to more than a single layer of perceptrons is that because of the discontinuous binary threshold, a small change in a weight in one layer could cause a major disturbance for the weights in the next.

By ''relaxing'' from a binary to a continuous, analog threshold, then, it is possible to slowly change weights in multiple levels without causing any major disturbances. This is at the basis of the back-propagation technique.

---

[18] Ackley, et. al. (1985, p. 152).

Given a set of inputs, x_i, a set of weights, w_i and a threshold, θ, a threshold logic unit will return 1 if:

$$(\sum_i x_i w_i) - \theta > 0$$

and 0 otherwise. The units used by the back-propagation procedure return:

$$\frac{1}{1+e^{\theta - \sum_i x_i w_i}}$$

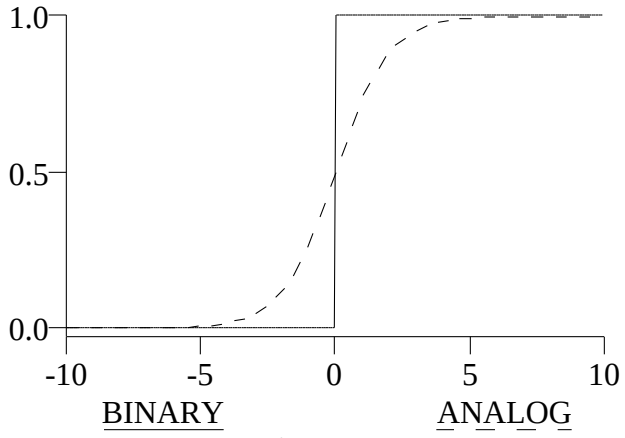


Figure 4.

Graphs of Binary versus Analog thresholding functions. The horizontal axis represents the linear combinations of inputs and weights, and the vertical axis shows the output function.

Graphs of these two functions are depicted in figure 4. It can be seen that for the analog case, small changes in the input (by slowly changing weights) cause correspondingly small changes in the output.

Back-propagation has been used quite successfully. (Sejnowski & Rosenberg, 1986) reported a text-to-speech program which was trained from phonetic data, and (Hinton, 1986) showed that, under proper constraints, back-propagation can develop semantically interpretable ‘‘hidden’’ features.

In fact, back-propagation is so widely being used today, that it is threatening to become a subfield of its own. One of the major foci of current connectionism is the application of back-propagation to diverse areas such as sonar (Gorman & Sejnowski, 1988), speech (Elman & Stork, 1987), machine translation (Allen, 1987), and the invention and investigation of numerous tweaks and twiddles to the algorithm (Cater, 1987; Dahl, 1987; Stornetta & Huberman, 1987).

Whether these new approaches to learning in multiple layers is more than a local maximum in the hill-climbing exercise known as science, is for future history to judge.

4.5. Facing the Future: Problems and Prognostications

Because of the problems to be described below, I cannot say, with conviction, that connectionism will solve major problems for Artificial Intelligence in the near future. I do not believe that the current intense military and industrial interest in neural networks will pay off on a grander scale than did the earlier commercialization of expert systems.

I do believe, however, that connectionism will eventually make a great contribution to AI, given the chance. Its own problems need to be solved first.

4.5.1. Problems for Connectionism

Despite the many well-known promises of connectionism, including massively parallel processing, machine learning, and graceful degradation, there are many limitations as well, which derive from naive applications of paradigmatic constraints derived from what is almost known about networks of real neurons. Many of these problems only arise when connectionism is applied to higher-level cognitive functions such as natural language processing and reasoning. These problems have been described in various ways, including: recursion, variable-binding, and cross-talk, but they seem to be just variations on older problems, for which entire fields of research have been established.

4.5.1.1. Generative Capacity

Despite the promises of connectionism, the paradigmatic assumptions lead to language processing models which are strictly finite-state. Several parsers have been built which parse context-free grammars of bounded length -- i.e. regular grammars. The term ‘‘generative capacity’’ is due to Chomsky, who used it as a measure of the power (capacity) of particular classes of formal grammars to generate natural language sentences; regular grammars are the weakest in this respect.

For example, as as adjunct to his model for word-sense disambiguation, Cottrell (1985) proposed a fixed-structure local connectionist model for length-bounded syntactic processing.

In a well-circulated report, Fanty (1985) describes the automatic construction a connectionist network which parses a context-free grammar. Essentially a time-for-space tradeoff, his system can parse bounded-length sentences, when presented all lexical items at once. The number of units needed for his network to parse sentences of length n rises as $O(n^3)$.

Selman (1985) also reports an automatic construction for networks which can parse bounded-length context-free grammars. His system is stochastic, and based on the Boltzmann Machine notions of (Ackley et al., 1985). Again we have a machine for sentences of bounded length. Another feature of Selman’s system is that the connectionist constraint of limited processing cycles is ignored and a parse may take several thousand cycles of annealing.

And even the newer crop of research in this area suffers from the same fixed-width problem (Allen, 1987; Hanson & Kegl, 1987; McClelland & Kawamoto, 1986).

4.5.1.2. Representational Adequacy

Closely related to the problem of generative capacity is the problem of representational adequacy. One must be careful that a model being proposed can actually represent the elements of the domain being modeled. One of the major attacks on connectionism has been on the inadequacy of its representations, especially on their lack of compositionality (Fodor & Pylyshyn, 1988). In feature-based distributed representations, such as the one used by (Kawamoto, 1985), if the entire feature system is needed to represent a single element, then attempting to represent a structure involving those elements cannot be managed in the same system. For example, if all the features are needed to represent a *Nurse*, and all the features are needed to represent an *Elephant*, then the attempt to represent a *Nurse riding an elephant* will come out either as a *white elephant* or a rather *large nurse with four legs*.

One obvious solution to this problem of superimposition versus concatenation involves using separate ‘‘pools’’ of units to represent elements of propositional triples, such as Agent, Action, and Object. In each pool would reside a distributed representation filling these roles such as ‘‘Nurse’’, ‘‘Riding’’, and ‘‘Elephant’’. Because of the dichotomy between the representation of a structure (by concatenation) and the representation of an element of the structure (by features), this type of system cannot represent recursive propositions such as "John saw the nurse riding an elephant".

Finally, parallel representations of sequences which use implicit sequential coding (such as Rumelhart and McClelland (1986) used in their perceptron model for learning the past tenses of verbs) have limits representing repetitive constituents. So a system, for example, which represented words as as collections of letter-triples, would not be able to represent words with duplicate triples such as *Banana*.

4.5.1.3. Task Control

A final problem is that many neural models use every ‘‘allowable’’ device they have to do a single task. This leaves no facility for changing tasks, or even changing the size of tasks, except massive duplication and modification of resources. For example, in the past-tense model (Rumelhart & McClelland, 1986), there is no obvious means to conjugate from, say, past to present tense, without another 200,000 weights. In the Travelling Salesman network (Hopfield & Tank, 1985), there is no way to add a city to the problem without configuring an entire new network.

4.5.2. Predicting the future

The existence and recognition of these problems is slowly causing a change in the direction of near-term connectionist research. There are many ongoing efforts now on more serial approaches to recognition and generation problems (Elman, 1988; Gasser & Dyer, 1988; Jordan, 1986; Pollack, 1987), which may help overcome the problem of massive duplication in dealing with time. There is also research in progress along the lines of Hinton’s (unpublished) proposal for reduced descriptions as a way out of the superposition/concatenation difficulty for distributed representations. For example (Pollack, 1988) demonstrates a reconstructive distributed memory for variable sized trees, and (Dyer et al., 1988) show a network construction for representing simple semantic networks as labelled directed graphs.

As problems in capacity, representation, and control get solved, we may expect a new blooming of connectionist applications in areas currently dominated by traditional symbolic processing.

I believe that connectionism may lead to an implementational redefinition of the notion of ‘‘symbol’’. In AI, symbols have no internal structure and thus mean very little, they are just used as names for, or pointers to, larger structures of symbols, which are reasoned with (slowly). The essential difference between the early neural network research and modern connectionism is that AI has happened in-between them. Because modern connectionism really does focus on representations, there is a possibility that a new kind of symbol might emerge from connectionism. For example, a reduced representation of some structure into a distributed pattern could be considered such a symbol, given that it can ‘‘point’’ to a larger structure through a reconstruction algorithm. Such ‘‘supersymbols’’ (as opposed to subsymbols (Smolensky, To appear)) may have an advantage over AI style token-symbols, in that they possess internal structure which can be reasoned about.

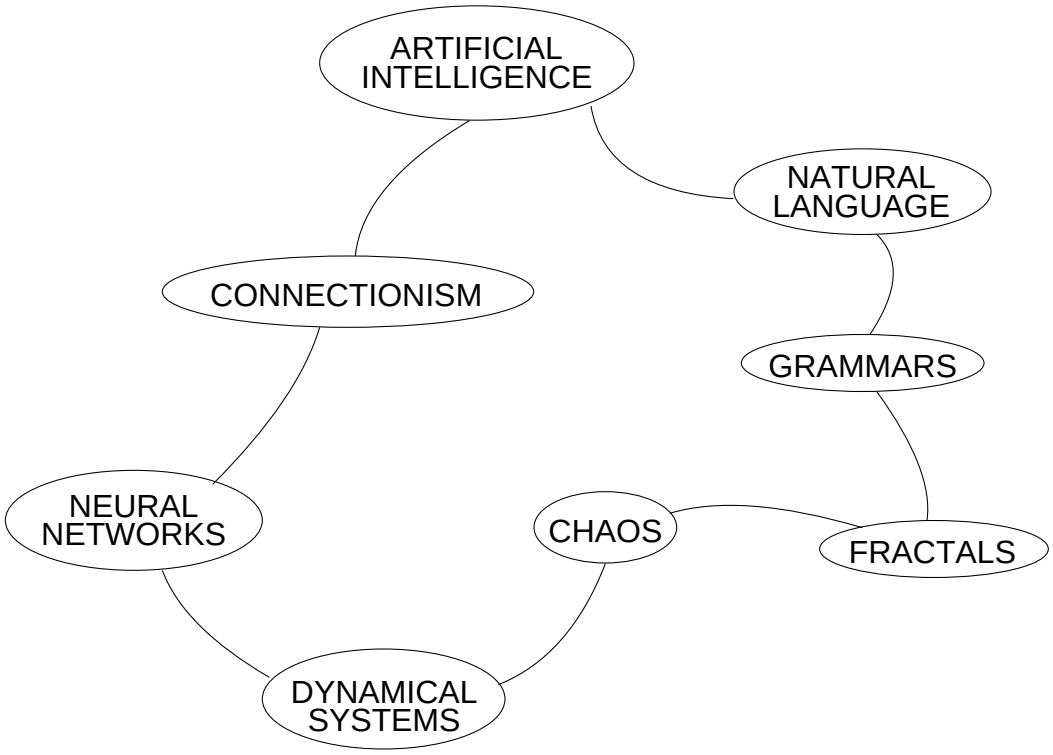


Figure 5.

Research areas which almost communicate.

Finally, I wish to make quite a far-fetched prediction, which is that Connectionism will sweep AI into the current revolution of thought in the physical and biological sciences (Crutchfield et al., 1986; Gleick, 1987; Grebogi et al., 1987). Figure 5 shows a set of disciplines which are almost communicating today, and implies that the shortest path between AI and Chaos is quite long.

There has already been some intrusion of interest in chaos in the physics-based study of neural networks as dynamical systems. For example both (Huberman & Hogg, 1987) and (Kurten, 1987) show how phase-transitions occur in particular neural-like systems, and (Lapedes, 1988) demonstrate how a network trained to predict a simple iterated function would follow that function’s bifurcations into chaos. However, these efforts are strictly bottom-up and it is still difficult to see how chaos has anything to do with connectionism, let alone AI.

Taking a more top-down approach, consider several problems which have been frustrating for some time. One problem is how to get infinite generative capacity into a system with finite resources (i.e., the competence/performance distinction). Another is the question of reconstructive memory, which has only been crudely approximated by AI systems (Dyer, 1983). Yet another is the symbol-grounding problem, which is how to get a symbolic system to touch ground in real-world perception and action, when all systems seem to bottom out at an *a priori* set of semantic primitives.

My suspicion is that many of these problems stem from a tacit acceptance, by both AI researchers and connectionists, of ‘‘aristotelian’’ notions of knowledge representation, which stop at terms, features, or relations. Just as Mandelbrot claims to have replaced the ideal integer-dimensional euclidean geometry

with a more natural fractional dimensional (fractal) geometry (Mandelbrot, 1982), so may we have ultimately to create a non-aristotelian representational base.

I have no concrete idea on what such a substrate would look like, but consider something like the Mandelbrot set as the basis for a reconstructive memory. Most everyone has seen glossy pictures of the colorful shapes that recurrently appear as the location and scale are changed. Imagine an ‘‘inverse’’ function, which, given an underspecified picture, quickly returns a ‘‘pointer’’ to a location and scale in the set. Reconstructing an image from the pointer fills in the details of the picture in a way consistent with the underlying self-similarity inherent in the memory. Given that all the representations to be ‘‘stored’’ are very similar to what appears in the set, the ultimate effect is to have a look-up table for an infinite set of similar representations which incurs no memory cost for its contents. Only the pointers and the reconstruction function need to be stored.

While it is not be currently feasible, I think that approaches like this to reconstructive memory may also engender systematic solutions to the other problems, of finitely regressive representations which bottom out at perception rather than at primitives, and which give the appearance of infinite generative capacity.

5. Conclusion

Like many systems considered historically, connectionism seems to have a cyclical nature. It may well be that the current interest dies quite suddenly due to the appearance of another critical tour-de-force such as *Perceptrons*, or, a major accident, say, in a nuclear power plant controlled by neural networks. On the other hand, some feel that AI is entering a retrenchment phase, after the business losses recently suffered by its high-profile corporate entities and the changing of the guard at DARPA. Given that it doesn't all go bust, I predict that the current limitations of connectionism will be understood and/or overcome shortly, and that, within 10 years, ‘‘connectionist fractal semantics’’ will be a booming field.

5.1. References

Ackley, D. H., Hinton, G. E. & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann Machines. *Cognitive Science, 9*, 147-169.

Allen, R. (1987). Several Studies on Natural Language and Back Propagation. In *Institute of Electrical and Electronics Engineers First International Conference on Neural Networks*. San Diego, II-335-342.

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.

Ashby, W. R. (1960). *Design for a Brain: The origin of adaptive behaviour (Second Edition)*. New York: John Wiley & Sons.

Barto, A. G., Anderson, C. W. & Sutton, R. S. (1982). Synthesis of Nonlinear Control Surfaces by a layered Associative Search Network. *Biological Cybernetics, 43*, 175-185.

Barto, A. G. (1985). Learning by statistical cooperation of self-interested neuron-like computing elements. *Human Neurobiology, 4*, 229-256.

Cater, J. P. (1987). Successfully using peak learning rates of 10 (and greater) in back-propagation networks with the heuristic learning algorithm. In *Institute of Electrical and Electronics Engineers First International Conference on Neural Networks*. San Diego, II-645-652.

Cottrell, G. W. (1985). Connectionist Parsing. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*. Irvine, CA.

Cottrell, G. W. (1985). A Connectionist Approach to Word-Sense Disambiguation. TR154, Rochester: University of Rochester, Computer Science Department.

Crutchfield, J. P., Farmer, J. D., Packard, N, H. & Shaw, R. S. (1986). Chaos. *Scientific American, 255*, 46-57.

Cun, Y. Le (1985). A Learning Scheme for Asymmetric Threshold Networks. In *Proceedings of Cognitiva 85*. Paris, 599-604.

Dahl, E. D. (1987). Accelerated learning using the generalized delta rule. In *Institute of Electrical and Electronics Engineers First International Conference on Neural Networks*. San Diego, II-523-530.

Dreyfus, H. L. & Dreyfus, S. E. (1988). Making a Mind versus Modeling the Brain: Artificial Intelligence Again at the Crossroads. *Daedalus, 117*.

Dyer, M. G. (1983). *In Depth Understanding*. Cambridge: MIT Press.

Dyer, M. G., Flowers, M. & Wang, Y. A. (1988). Weight Matrix = Pattern of Activation: Encoding Semantic Networks as Distributed Representations in DUAL, a PDP architecture. UCLA-Artificial Intelligence-88-5, Los Angeles: Artificial Intelligence Laboratory, UCLA.

Elman, J. & Stork, D. (1987). Session on Speech Recognition and Synthesis. In *Institute of Electrical and Electronics Engineers First International Conference on Neural Networks*. San Diego, IV-381-504.

Elman, J. L. (1988). Finding Structure in Time. Report 8801, San Diego: Center for Research in Language, UCSD.

Fanty, M. (1985). Context-free parsing in Connectionist Networks. TR174, Rochester, N.Y.: University of Rochester, Computer Science Department.

Fodor, J. & Pylyshyn, A. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition, 28*, 3-71.

Gasser, M. & Dyer, M.G. (1988). Sequencing in a Connectionist Model of Language Processing. In *Proceedings of the 12th International Conference on Computational Linguistics*. Budapest.

Gleick, J. (1987). *Chaos: Making a new science.* New York: Viking.

Gorman, R. P. & Sejnowski, T. J. (1988). Analysis of hidden units in a layered nework trained to classify sonar targets. *Neural Networks, 1*, 75-90.

Grebogi, C., Ott, E. & Yorke, J. A. (1987). Chaos, Strange Attractors, and Fractal Basin Boundaries in Nonlinear Dynamics. *Science, 238*, 632-638.

Grossberg, S. (1987). Competitive Learning: From Interactive Activation to Adaptive Resonance. *Cognitive Science, 11*, 23-63.

Hanson, S. J. & Kegl, J. (1987). PARSNIP: A connectionist network that learns natural language grammar from exposure to natural language sentences. In *Proceedings of the Ninth Conference of the Cognitive Science Society.* Seattle, 106-119.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. New York: John Wiley & Sons.

Hinton, G. E. (1986). Learning Distributed Representations of Concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society.* Amherst, MA, 1-12.

Hopfield, J. J. (1982). Neural Networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA, 79*, 2554-2558.

Hopfield, J. J. & Tank, D. W. (1985). 'Neural' computation of decisions in optimization problems. *Biological Cybernetics, 52*, 141-152.

Huberman, B. A. & Hogg, T. (1987). Phase Transitions in Artificial Intelligence Systems. *Artificial Intelligence, 33*, 155-172.

Jordan, M. I. (1986). Serial Order: A Parallel Distributed Processing Approach. ICS report 8608, La Jolla: Institute for Cognitive Science, UCSD.

Kawamoto, A. H. (1985). Dynamic Processes in the (Re)Solution of Lexical Ambiguity. Doctoral Dissertation, Providence: Department of Psychology, Brown University.

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*, 671-680.

Klopf, A. H. (1982). *The Hedonistic Neuron*. Washington, D.C.: Hemisphere Publishing Corporation.

Kohonen, T. (1977). *Associative Memory: A Systems-Theoretical Approach.* Berlin: Springer-Verlag.

Kohonen, T., Oja, E. & Lehtio, P. (1981). Storage and Processing of Information in Distributed Associative Memory Systems. In G. E. Hinton & J. A. Anderson, (Eds.), *Parallel models of associative memory.* Hillsdale: Lawrence Erlbaum Associates.

Kurten, K. E. (1987). Phase transitions in quasirandom neural networks. In *Institute of Electrical and Electronics Engineers First International Conference on Neural Networks*. San Diego, II-197-20.

Lenat, D. B. (1977). The Ubiquity of Discovery. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence.* Cambridge, MA, 1093-1105.

Mandelbrot, B. (1982). *The Fractal Geometry of Nature*. San Francisco: Freeman.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of the effect of context in perception: Part 1. An account of basic findings. *Psychology Review, 88*, 375-407.

McClelland, J. & Kawamoto, A. (1986). Mechanisms of Sentence Processing: Assigning Roles to Constituents. In J. L. McClelland, D. E. Rumelhart & the PDP research Group, (Eds.), *Parallel Distributed Processing: Experiments in the Microstructure of Cognition*, Vol. 2. Cambridge: MIT Press.

McCullogh, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115-133.

Minsky, M. & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Minsky, M. (1986). *The Society of Mind*. New York: Simon & Schuster.

Norman, D. A & Rumelhart, D. E. (1975). *Explorations in Cognition*. San Francisco: W. H. Freeman & Co..

Parker, D. B. (1985). Learning-Logic. Technical Report-47, Cambridge: MIT Center for Computational Research in Economics and Management Science.

Pollack, J. B. (1987). Cascaded Back Propagation on Dynamic Connectionist Networks. In *Proceedings of the Ninth Conference of the Cognitive Science Society.* Seattle, 391-404.

Pollack, J. B. (1988). Recursive Auto-Associative Memory: Devising Compositional Distributed Representations. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society.* Montreal.

Rosenblatt, F. (1959). Two theorems of statistical separability in the perceptron. In *Mechanization of Thought Processes*, Vol. 1. London: Her Majesty's Stationary Office.

Rosenblatt, F. (1962). *Principles of Neurodynamics*. New York: Spartan.

Rumelhart, D. E. & McClelland, J. L. (1982). An interactive activation model of the effect of context in perception: Part 2 The contextual enhancement effect and some tests and extensions of the model. *Psychology Review, 89*, 60-94.

Rumelhart, D. E. & Zipser, D. (1986). Feature Discovery by Competitive Learning. In D. E. Rumelhart, J. L. McClelland & the PDP research Group, (Eds.), *Parallel Distributed Processing: Experiments in the Microstructure of Cognition*, Vol. 1. Cambridge: MIT Press.

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: Experiments in the Microstructure of Cognition*, Vol. 2. Cambridge: MIT Press.

Sabbah, D. (1982). A Connectionist Approach to Visual Recognition. TR107: University of Rochester, Computer Science Department.

Sampson, G. (1986). A stochastic approach to parsing. In *COLING*. Bonn.

Sejnowski, T. J. & Rosenberg, C. R. (1986). NETtalk: A parallel network that learns to read aloud. JHU/EECS-86/01: The Johns Hopkins University, Electrical Engineering and Computer Science Department.

Selman, B. (1985). Rule-Based Processing in a Connectionist System for Natural Language Understanding. CSRI-168, Toronto, Canada: University of Toronto, Computer Systems Research Institute.

Shastri, L. (1988). *Semantic Nets: An evidential formalization and its connectionist realization*. Los Altos, CA: Morgan Kaufmann.

Smolensky, P. (1986). Information Processing in Dynamical Systems: Foundations of Harmony Theory. In D. E. Rumelhart, J. L. McClelland & the PDP research Group, (Eds.), *Parallel Distributed Processing: Experiments in the Microstructure of Cognition*, Vol. 1. Cambridge: MIT Press.

Smolensky, P. (To appear). On the proper treatment of Connectionism. In *Behavioral and Brain Sciences*. .

Stornetta, W. S. & Huberman, B. A. (1987). An Improved three-layer back propagation algorithm. In *Institute of Electrical and Electronics Engineers First International Conference on Neural Networks*. San Diego, II-637-644.

Touretzky, D. S. & Hinton, G. E. (1985). Symbols among the neurons: details of a connectionist inference architecture. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*. Los Angeles, CA.

Willshaw, D. J. (1981). Holography, Associative Memory, and Inductive Generalization. In G. E. Hinton & J. A. Anderson, (Eds.), *Parallel models of associative memory*. Hillsdale: Lawrence Erlbaum Associates.