Neural Networks in Cognitive Science

Jeff Yoshimi

September 7, 2011

Chapter 2: Historical and Philosophical Context

In this chapter I give a brief historical overview of connectionism, then describe philosophical differences between connectionism and its main rival (at least until recently): symbolic artificial intelligence.

1 A brief history of connectionism

1.1 Ancient roots

Cognitive science, the interdisciplinary study of mind, has ancient roots. In the Western tradition, Plato, Aristotle, and other Greek philosophers had an interest in the structure of the human mindor "soul" in relation to physical processes in the body. Plato and Aristotle both described the soul as a set of interacting faculties (in Plato: reason, spirit, and appetite), and both speculated about its physical basis. They disagreed about whether the brain or heart is the physical basis of the soul (Aristotle thought the brain just cooled the blood), but by the end of the Classical period the brain was generally thought to be the physical basis of the mind.

1.2 The Medieval Period

The question of how the mind arises from the brain was subsequently taken up by medieval thinkers, who believed cognition was based on the play of "spirits" or vapors in the ventricles of the brain. Spirits originating in the senses were combined in the "common sense" and then purified, and mixed in higher ventricles. A typical diagram from the period is shown in figure 1. Today the ventricles are believed to be shock-absorbers and chemical reservoirs, and are not thought to play a central functional role in cognition. However, the idea that sensory inputs to the brain are combined and refined in various ways persists in connectionist models.

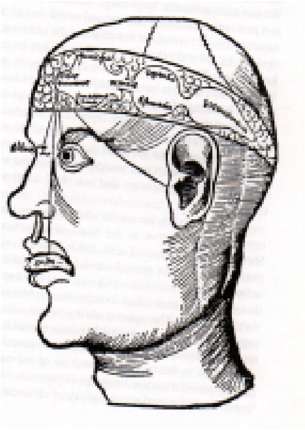

Figure 1: A medieval diagram which shows how spirits were thought to flow in the ventricles of the brain. From Grüsser, 1990.

1.3 The Enlightenment

During the Enlightenment, various thinkers proposed that connections between ideas in the mind are based on connections between fibers in the brain.[1] For example, in the 1700's Locke famously claimed that ideas in the mind result from associations between simple sensory ideas: for example, a percept of an apple is composed out simple sensations corresponding to its color, shape, smell, and taste. Several thinkers after Locke, including Hartley and later, Bain, thought that Locke's theory could be explained by laws describing connections between neurons in the brain. A diagram from Bain is shown in figure 2.[2]

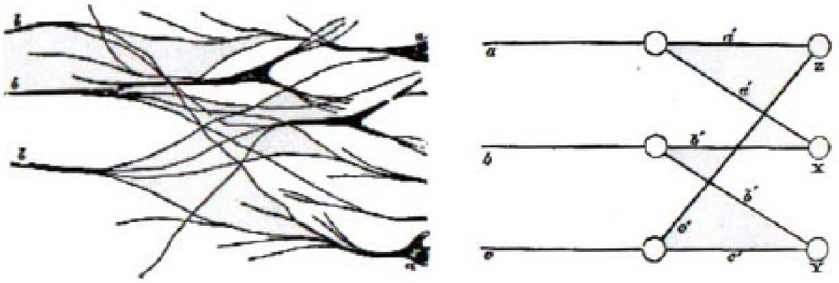

Figure 2: From Bain, 1873

[1] Neurons had not yet been identified as distinct structures.

[2] For more on this period of history, see Sutton (1998).

1.4 Twentieth Century

In the early part of the 20th century a movement which anticipated connectionism was cybernetics, “the science of control and communication in the animal and the machine” (Wiener, 16). Early cybernetics was fueled by industry and war. In the late 1930s the petroleum industry developed central control systems to maintain refinery towers. These feedback systems were pressed into military service during WWII, when engineers were asked to design more efficient mechanisms for controlling anti-aircraft guns. Before long, these engineering techniques were applied to the study of mind. It was realized that feedback based circuits could coordinate complex movement, both in engineered systems and in the brain.

In the 1940’s Warren McCulloch (a neurophysiologist affiliated with cybernetics) and Walter Pitts (a logician) famously showed how neuron-like elements, operating in parallel, could perform all the logical operations performed by computers. This in turn implies that whatever can be done on a computer can, in principle, be done using neurons. A diagram from one of McCulloch’s papers is shown in figure 3.

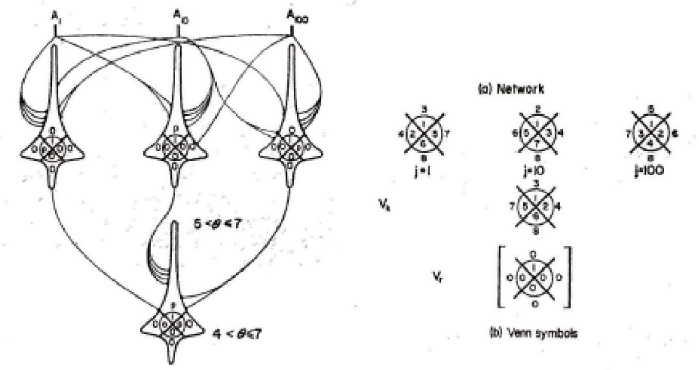

Figure 3: From McCulloch, 1960.

Another important figure in this period was Donald Hebb, sometimes called the father of connectionism. Hebb was a psychologist who formulated a famous learning rule for weights, the “Hebb rule” (“neurons that fire together, wire together”), and who explicitly thought about the operation of the brain in terms of networks of connected neurons. In particular, he formulated the concept of a “cell assembly”, a groups of neurons that become associated over time and thereafter tend to reverberate in response to a stimulus (figure 4 shows one of Hebb’s own diagrams of a cell asembly).

Connectionism in its present form developed in the context of the early “cognitive revolution”, which in turn led to what is today called “cognitive science”. The cognitive revolution is sometimes traced to one of several conferences in the 1950s, including a conference held at MIT in 1956, the Symposium on Informa-

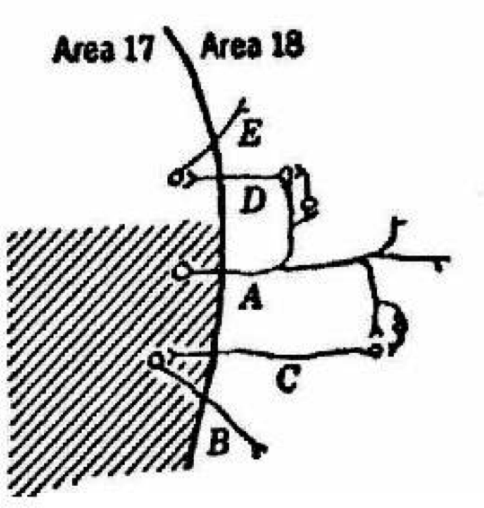

Figure 4: A Hebbian cell assembly.

tion Theory:

> Three talks in particular, Miller's 'The magical number seven', Chomsky's 'Three models of language', and Newell and Simon's 'Logic theory machine', have been singled out as instrumental in seeding the cognitive science movement. Following these talks, a perception began to emerge that "human experimental psychology, theoretical linguistics, and computer simulations of cognitive processes were all pieces of a larger whole" (Medler; the embedded quote is from Miller 1979).

Conferences like this were in part a response to behaviorism, according to which psychologists should focus on observable sensory inputs and motor responses, treating the "mind" as a black box between the inputs and outputs, that could be ignored. The new cognitive scientists wanted to break open that black box and look inside: they wanted to understand what kind of processing occurs between sensory input and motor output. The big idea, the idea that got everyone excited, was that inside the mind there is an information processing system, one that can be studied using computer simulations.

This new breed of inter-disciplinary cognitive scientists would, however, split, according to how they thought the mind processes information. Some thought (and some still think) that the mind processes information using symbols and rules, like a regular computer does. This remains a dominant tendency in cognitive science, and it goes by several names: "functionalism", "Symbolic AI," and "Classical AI", among others. The other camp, the "connectionists", thought the mind processes information more like the brain does. Today there are other camps as well, and some who mix ideas from both symbolic AI and connectionism. We will go into this difference in section 2.

In the 1950's and 1960's what would today be recognized as neural network theory really began. Rosenblatt, Widrow, Hoff, and others showed how simple networks could learn to recognize patterns using rules they derived mathematically (more on these rules in a later chapter). In this period some of the first neural networks were implemented in hardware–for example, Widrow and Hoff created a neural network architecture called "Adaline". In the hardware implementation shown in figure 5, the knobs are input neurons, the toggle switches are synapses (which can only be on or off), and the dial shows the activation of an output neuron.

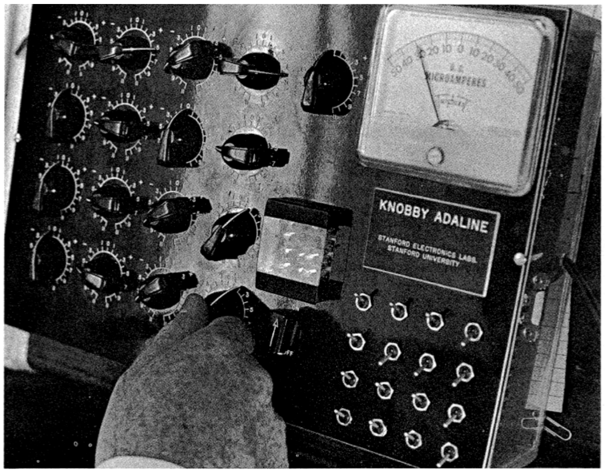

Figure 5: Adaline. A 2-layer network, with 12 input nodes, and one output node

Another important figure in the 1950's and 1960's was the psychologist Oliver Selfridge, who pioneered the idea that psychological processes can be broken down in to smaller process. His Pandemonium theory described the mind as a collection of sub-processes or "demons", each of which takes care of one specific aspect of a task. For example, figure 6 shows how Selfridge thought of the process of perceiving the letter "B". An image arrives at the eye, line demons detect lines in various orientations, and those demons send messages to angle demons who detect angles, and the process continues through a network of demons until a decision demon says "B!"

In the late 1960's and early 1970's, the symbolic AI model of cognition was dominant. Moreover, simple neural networks like the ones studied by Rosenblatt (simple 2-layer networks) were shown to suffer certain fundamental limitations. For these and other reasons, neural networks were relatively unpopular in this period. In fact, these have been called the "dark ages" of connectionism, though

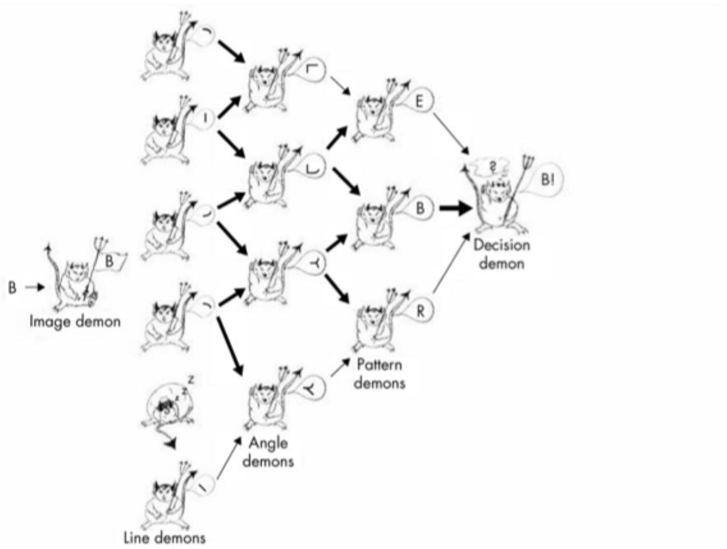

Figure 6: Selfridge's pandemonium model.

it was only a decade or so, and lots of good stuff happened during that decade.[3]

Connectionism came out of its dark decade and enjoyed a resurgence in the 1980s, for several reasons, including the discovery of the backpropagation algorithm, which addressed the problems that affected earlier networks, and the emergence of a number of limitations associated with classical AI. The resurgence was also helped by the publication of a major two-volume work in the period, *Parallel Distributed Processing: Adventures in the Microstructure of Cognition*, in 1986, by David Rumelhart, James MClelland, and the "PDP research group" (a group of researchers, many of whom were at UC San Diego.) This publication brought connectionist networks back to the forefront, by clearly articulating the connectionist standpoint, showcasing a number of models of various aspects of cognition, and clarifying how connectionist networks differ from symbolic AI models.

2 Symbolic AI vs. Connectionism

As noted above, soon after the emergence of early cognitive science, a split developed between those who de-emphasized the brain (symbolic AI researchers), focusing instead on the AI / computer model of mind, and those (connectionists) who emphasized brain-like aspects of cognitive processing. Let us consider the contrast in more detail, by considering three assumptions of each approach.

Three basic assumptions of symbolic AI are that:

[3] For more on the work done in this period see PDP vol 1, ch. 1.

> 1) Neuroscience is not important for the study of cognition. According to this view, what matters to cognitive science is the formal structure of cognition, not "implementation" details concerning the brain. As it was sometimes put, "the mind is to the brain as software is to hardware", or " the mind is software running in the brain". Compare a computer program like Photoshop. Someone who wanted to know how Photoshop works would care about its high-level source code, but not about the physical transistors in the computer. On this view, what matters is high-level formal structure, not low-level neural details.
>
> 2) Mental representations are symbol structures with a linguistic form. On this view, the mind is a computer program written in a kind of language, the "language of thought" (sometimes called "mentalese"). Mental processes correspond to operations on these symbol structures. This implies that the basic units of thinking are discrete linguistic units which get pieced together into more complex representations: for example, my thought that the computer is in front of me is a sentence in my mind, made out of discrete parts like "computer" and "front".
>
> 3) Thinking involves applying explicit rules to symbolic structures. For example, two minutes from now I can think "that computer *was* in front of me" by changing the tense of the verb "is" using a rule-based transformation "is" > "was". So while the elements of thought are symbol structures, the process of thinking consists in applying explicit rules to those structures.[4]

Connectionism has been put forward as an alternative to symbolic AI. Three fundamental assumptions of connectionism, which contrast with those of symbolic AI, are as follows:

> 1) Neuroscience is essential to studying cognition. To understand the mind we need to understand the brain, and should use brain-like models.
>
> 2) Mental representations correspond to patterns of activity across many nodes. Mental representations correspond to patterns of activity over millions of nodes, rather than language-like symbol structures.

[4] This is a somewhat simplistic way of describing Classical AI, and some who endorse it, or something like it, would not say things this way. My purpose is to set up a first-pass, strong contrast.

> 3) Thinking involves transforming patterns of activity by patterns of connections. On this view, thinking involves the changes that occur as neural activity passes through complex webs of synaptic connections.

So, whereas for symbolic AI the mind is a like a computer, programmed with rules which operate on symbolically structured representations, for connectionism the mind is like the brain, transforming patterns of neural activity via complex webs of synaptic connections.

This division is not as sharp as it used to be, and many today use both kinds of model, or hybrid models (more below).

Having seen the general division between symbolic AI and connectionism, let us describe special features of neural networks that have been taken to demonstrate their advantages relative to classical symbol systems.

2.1 Neural Networks Operate in Parallel

Whereas digital computers normally do things one at a time, neural networks do a lot of things at the same time. Digital computers perform computations in serial, while neural networks perform computations in parallel. To see the difference, consider a simple problem: finding which of ten cups has a bean under it. A serial approach would lift each cup up, one at a time, until the bean was found. A parallel approach would lift all ten cups up at once.

Parallel computation is clearly faster. So, why not do everything in parallel? The answer, roughly, is that the whole theory of digital computation assumes a core level of serial processing. To implement a basic algorithm assumes that things happen in a well-ordered fashion. Techniques for parallel computation exist, and are emerging as an important area of computer science, but even in those cases tasks are broken down into sub-tasks each of which can be run in serial. Moreover, serial computers are fast these days. For example, the computer I'm writing this on performs several billion operations per second. With numbers like those we can tolerate the processing slow-down due to serial processing.

On the other hand, massively parallel processors like the brain are messy, they make mistakes. They are not as good as computers at performing, e.g., complex mathematical computations or looking up names in an index. And they are made of relatively slow materials. Neurons fire at most 200 times per second. They are orders of magnitude slower then the transistors on this computer. In order to solve problems in a reasonable time frame the brain operates in parallel: every neuron is always doing its own thing. The trade-off is in accuracy: the brain is a kind of messy computer, which doesn't always do the same thing twice. But given the wet, biological stuff we're made of, it works well enough.

Even though neural networks operate in parallel, many aspects of thinking are serial. As Rummelhart , McClelland, and Hinton point out:

> The process of human cognition, examined on a time scale of seconds and minutes, has a distinctly sequential character to it. Ideas

> come, seem promising, and then are rejected; leads in the solution to a problem are taken up, then abandoned and replaced with new ideas... Clearly any useful description of the overall organization of this sequential flow of thought will necessarily describe a sequence of states (p. 12).

Does this mean neural networks are really modeling a sequential process? Not really. Neural networks model the "microstructure" of cognition, what is sometimes called the "sub-symbolic" level of mental processing. Even if it is true that the overall behavior of our brain–patterns of firing across millions of neurons–flows from macro-pattern to macro-pattern sequentially, these high level processes are somehow grounded in low level processes involving thousands or millions of parallel computations.

2.2 Neural Networks Gracefully Degrade

Classical, digital computers are "brittle," in the sense that if a single component is lost it will probably stop functioning properly. Knock out the microprocessor in your computer, or snip a few wires, and the whole thing will most likely stop functioning altogether.

Neural networks, by contrast, gracefully degrade (we discussed this briefly in chapter 1). If you lose a few neurons and /or synapses, there is a good chance that the whole system will continue to function well. Of course, if you lose enough neurons and synapses it will show, but the damage in performance is generally proportional to the damage to the network. Neural networks degrade "gracefully." This is related to the fact that they operate in parallel rather than serial. While a serial computer needs to have every component linked up properly to work, a neural network has lots of redundant wiring which can compensate for damage.

2.3 Neural Networks are Tolerant of Noisy Inputs

Digital computers don't like noisy input: they respond only to clean, precise inputs. Anyone who has worked with computers has some understanding of this. To get through a company's "phone tree" you have to enter just the right sequence of numbers–no mistakes allowed! Or, suppose you are looking me up in a database. If you enter "Joshimi" instead of "Yoshimi" you won't find me. You must enter the exact right input to get the right response.[5]

Neural networks, on the other hand, do quite well with noisy inputs. Just look at our brains. Show me ten roses, and the exact pattern of stimulation on my retina will differ. In fact, show me the same rose ten times, and there is sure to be noise in the pattern produced on my retina. But I see it as a rose every

[5] Of course Google and other computers can make good guesses, but then they are using technologies inspired by neural network ideas; they are approximating neural network like computation in a digital computer.

time. We will see lots of examples of neural networks which keep chugging along even though the inputs presented to them are noisy.

In a sense, this is the concept of graceful degradation, applied to inputs rather than processing components. Brains do well with noisy, "degraded" inputs, but digital computers generally don't.

2.4 Neural Networks use Distributed Representations

Mental representations (e.g. your knowledge of your grandmother) can be thought of in two ways: as being locally stored in one location in the brain, or as being distributed over many locations. A local representation scheme for the brain is sometimes called a "grandmother cell" doctrine, because it implies that there is just one neuron in your brain that represents your grandmother.

In the context of neural networks, we can say that an object P is locally represented by a neuron when activation of that neuron indicates the presence of P. For example, in figure 7, blue cheese is locally represented by the neuron labelled "Center 5". When that neuron is activated, the blue cheese is present (here "activation" means having a non-zero, positive activation value).

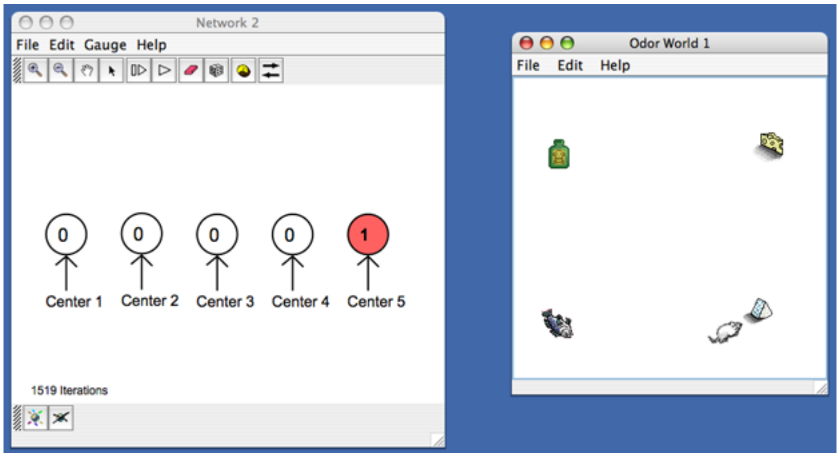

Figure 7: Localist representation

In contrast, we can say that an object has a distributed representation in a neural network, when a particular *pattern of activation* over a set of nodes indicates the presence of that object. In figure 8, the bottle of poison has a distributed representation. When the poison is present a specific pattern of activation $(.1, 1, .7, 0, .2)$ occurs across the whole set of nodes.

For the most part, distributed representations are what one finds in the brain. Generally speaking brain functions are distributed over many neurons.

Although it is harder to think about distributed representations than about local representations, we see in another chapter how these patterns can be visualized as points in a space. This in turn makes it easier to visualize their

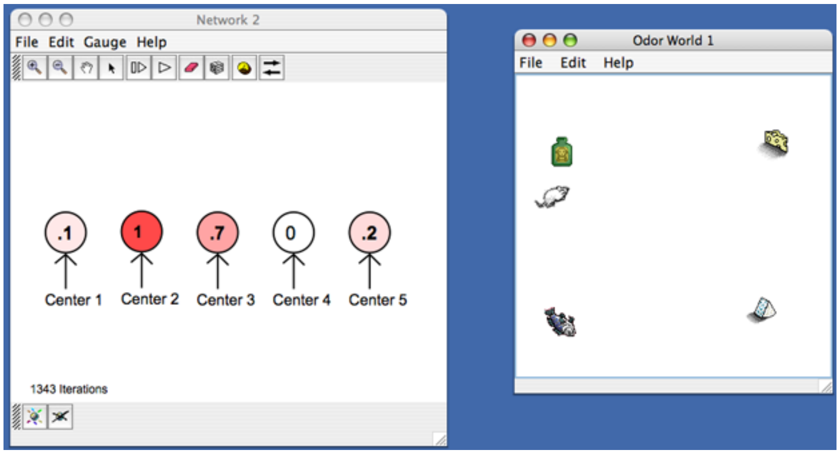

Figure 8: Distributed representation

relations to one another.

Some older types of neural network use only local representations (e.g. the IAC networks in chapter 1), and we will see that it is sometimes useful to use local representations. However, the problem with local representations is that you lose some of the virtues above, in particular graceful degradation. If there is just one unit whose activation represents my grandmother, then if I lose that neuron I lose my whole memory of my grandmother. But the empirical evidence suggests that losing a single neuron will not lead to a person's losing an entire memory. So, even if some artificial neural networks use localist schemes to illustrate certain concepts, biological neural networks don't seem to.

3 Middle / Hybrid Positions

The split between symbolic AI and connectionism continues to this day: for example, even today the role of neuroscience in cognitive science is a controversial topic. But though I have set up a strong split between AI and connectionism, there are many today who do both (or work with other types of models, e.g. Baysean models). In a way it becomes a matter of what level of analysis a cognitive science researcher works at. If someone is interested in low level perception, a connectionist approach might be better; if someone is interested in language processing, AI might be better. In the end I think most theorists and working scientists have an ecumenical attitude: choose whatever model works best to describe the data you have.

References and Suggested Readings

Bain, A. 1873. Mind and Body: The Theories of Their Relation. London: Henry S. King & Co.

Gardner. H. The Mind's New Science. Basic Books, New York, 1985.

Grsser, O. J. 1990. On the Seat of the Soul, Cerebral Localization Theories in Medieval Times and Later, in Brain-Perceptual Cognition. Proceedings of the 18th Gttingen Neurobiology Conference. Stuttgart: Verlag.

Medler, D. A Brief History of Connectionism.

McCulloch, W. 1960. The Reliability of Biological Systems in Self Organizing Systems. New York: Pergamon Press.

Rumelhart, D., McClelland, J, and the PDP Research Group. 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. See esp. ch. 1 and ch. 4.

Sutton, J. 1998. Philosophy and Memory Traces: Descartes to Connectionism. Cambridge: Cambridge University Press.

Wiener, N. 1965 (1948). Cybernetics. MIT Press.