Connectionism Reconsidered: Minds, Machines and Models

by

István S. N. Berkeley Ph.D.
(istvan@Louisiana.edu)

Philosophy and Cognitive Science,
The University of Louisiana at Lafayette

Abstract

In this paper the issue of drawing inferences about biological cognitive systems on the basis of connectionist simulations is addressed. In particular, the justification of inferences based on connectionist models trained using the backpropagation learning algorithm is examined. First it is noted that a justification commonly found in the philosophical literature is inapplicable. Then some general issues are raised about the relationships between models and biological systems. A way of conceiving the role of hidden units in connectionist networks is then introduced. This, in combination with an assumption about the way evolution goes about solving problems, is then used to suggest a means of justifying inferences about biological systems based on connectionist research.

Keywords: Connectionism, backpropogation, cognitive modeling, model validation.

Contacts: István S. N. Berkeley, Ph.D.
Philosophy and Cognitive Science,
The University of Louisiana at Lafayette
P.O. Drawer 43770
Lafayette
LA 70504-3770
USA
Tel: (337) 482-6807.

Connectionism Reconsidered: Minds, Machines and Models

The Appeal of Connectionism

Modeling with connectionist systems is now an integral part of Cognitive Science. One of the putatively appealing features of connectionist models is that they are supposed to be, in some sense, more biologically plausible, or brain-like than models which have their roots in what Haugeland (1985) terms the 'GOFAI' (short for 'Good Old Fashioned Artificial Intelligence') tradition. This justification of connectionism has found a strong hold in the literature, especially the philosophical literature (see Clark 1989: p. 4, Bechtel and Abrahamsen 1991: p. 17, Churchland 1989: p. 160, Dennett 1991: p. 239, Sterelny 1990: p. 175 and Cummins 1989: p. 155, for examples). However, over the years, it has become increasingly clear that this justification is deeply suspect, when applied to certain important sub-classes of connectionist system.

Consider for example, the case of the backpropagation learning procedure. Although this learning procedure has been widely deployed, it has been well known for some time, that the procedure is highly biologically implausible, for a variety of reasons (Grossberg 1987 and Quinlan 1991). This being the case, the justification for employing a connectionist model that employs the backpropagation learning procedure on a particular problem, cannot be based just upon putative biological plausibility.

Although some network systems *can* be justified on biological grounds, for example those which fall into the field often referred to as 'neural computing' (Churchland and Sejnowski 1992: p. 14), systems which employ backpropagation fall outside the scope of this justification. The justification for backpropagation models thus must come from other grounds. What then, are the reasons for believing that backpropagation trained models can tell us anything about biological cognition? Sketching an outline of an answer to this question will be the main purpose of this paper. To being with, it is worth briefly considering the relationship between models and minds in general.

Back to Basics

As a minimal condition, any proposed model of some aspect of biological cognition must have the appropriate input and output behaviors to model the relevant aspect of cognition. That is to say, the model had better be able to do more or less the same things as the biological system that it putatively models. However, this alone is not sufficient to justify a particular model as a basis to draw inferences about biological cognitive function. This is because there are many different ways to compute any particular function.

Consider the example of multiplication. Whilst most of us are taught in school to do multiplication using, what is known as the 'classical multiplication algorithm', pocket calculators go about multiplying two numbers in an entirely different way. Calculators use an algorithm known as 'multiplication *a la Russe*' (see Brassard and Bratley 1988: p. 2). Of course, there is no straightforward way for the user of a calculator to know that the

machine is calculating in a different manner, as both means of calculating give the same results. The reason that a calculator uses this algorithm is because it is much easier and simpler to implement in digital circuits. However, it would be a great error for a researcher to think that they could learn something about the human ability to multiply numbers by studying a calculator. This is the reason why the mere fact that a model can apparently duplicate input and output behaviors of some aspect of cognitive functioning is not sufficient to justify inferences about biological cognitive agents. However, such duplication of input and output behavior does constitute a necessary condition for such inferences to be drawn.

Pylyshyn (1984) draws a distinction that is helpful in this context. It is the distinction between computational systems which are strongly equivalent to biological systems, verses those which are weakly equivalent to such systems. Systems or models which are merely weakly equivalent to some aspect of biological cognitive functioning may have the same input and output behaviors as some biological system in the relevant respects, however they will go about producing this behavior in a different way. One consequence of this is that although the set of behaviors being studied may be the same, emergent behaviors (i.e. those behaviors that the model is not explicitly designed for, which come 'for free' so to speak) will most likely be different. So, to continue the example above, a calculator doing multiplication *a la Russe* may be weakly equivalent to a human being doing classical multiplication, but as noted, this does not provide much of a basis for inference about human cognitive functioning. When a model or system is strongly equivalent to some biological system, by contrast, the system or model not only has the same input and output behaviors, but also computes the function *in the same way*. That is to say, if one system is strongly equivalent to another system, then the two systems are computing exactly the same algorithm, in the same way. Moreover, a consequence of this will be that strongly equivalent systems will have the same emergent behaviors. It is only in the case of systems or models that are strongly equivalent to biological systems that adequately justified inferences about the biological systems can be drawn directly on the basis of the non-biological ones.

There are grounds for believing that the study of systems which are merely weakly equivalent may nonetheless be illuminating, in an indirect sense, about biological cognition. Dennett (1978: p. 113), for example has argued that it is possible to learn much of significance to psychology and epistemology on the basis of particular, though unrealistic (compared to natural cognitive systems) models. Dennett's contention is that such models can provide information about the general principles governing psychological and epistemological (i.e. cognitive) systems. As it is highly doubtful that many of the connectionist systems which have been trained using the backpropagation learning procedure manage to reach the standards required of strongly equivalent systems, it is this potential utility of connectionist systems, as weakly equivalent systems, which initially will be developed further here. In order to do this, it is worth pausing briefly to consider what a system trained using the backpropagation procedure has to accomplish in order to reach convergence.

Problems and Properties

When a connectionist network is trained using backpropagation, a number of input patterns are presented to the network, usually after the weights of the network have been randomized. For each pattern, the network will produce some response that is then compared to a desired response. This enables changes to be made to the weights between the layers of processing units in the network. The changes are made such that the network will respond more closely to the desired response the next time it receives the same input, or set of inputs. Assuming that a network successfully learns a problem, the network will at least produce the desired response when presented with every input pattern in the training set.

In order to learn a particular problem, a network has to find regularities in the input data that will enable it to produce the correct output for the problem being trained. The exact nature and kind of regularities will depend upon the precise problem being trained. However, for many interesting problems (those which are non-linearly separable), finding the necessary regularities requires some kind of recoding of the input information. In order for a network to be able to do this, it requires hidden processing units (Clark and Thornton, 1997). The role of the hidden units is to recode the inputs, so as to make the solution to the problem possible by the network. This enables the hidden units to be conceived of as devices that serve to detect input properties which are important to the solution.

It also turns out that for some tasks, we can know in advance, *a priori*, what some of the input properties a network will have to become sensitive to, for a particular problem set. Consider the case of a network that has to learn to distinguish valid from invalid instances of a set of simple arguments. If the training set contains a range of connectives, then the network would have to take into account the main connective of a problem, so as to be able to distinguish, for example, an invalid instance of a Modus Ponens inference, from an valid instance of a Disjunctive Syllogism inference. A network successfully trained on a problem of just this kind has been described by Bechtel and Abrahamsen (1991). Subsequent analysis of a network which was successfully trained upon Bechtel and Abrahamsen's problem set revealed that the network had indeed learned to become sensitive to exactly this input property (Berkeley *et al*. 1995).

If we know that hidden units function as input property detectors and, in some instances, we can even predict what the kinds of property they will have to detect in order to solve certain problems, then connectionist networks can be conceived of (in principle at least) as offering the means to *discover* sets of properties which, in combination, can solve particular problems. However, given the discussion above of the potential strong/weak equivalence relations which can hold between computational systems and biological ones, there may be legitimate grounds for wondering exactly why this conclusion should be taken as being of particular interest to researchers interested in cognition. After all, why would there be grounds for thinking that a network will find a solution to a particular problem that is strongly equivalent, rather than weakly equivalent?

On the one hand it has been argued that if we want to learn and make justified inferences about biological cognition from computational models, we really need to ensure that the models we draw inferences from are strongly equivalent to the biological cognitive systems of interest. On the other hand, it has been argued that we may be able to draw inferences about the general class of systems with certain apparently cognitive capacities, by studying only weakly equivalent systems. It has also been argued that connectionist networks, which are trained using the backpropagation training procedure, offer a means of determining the sets of input properties that are important to solving particular problems. Yet, as the number of algorithms for solving particular problems is in principle intractably large, it would seem likely that the properties discovered by the hidden units of individual backpropagation networks, could reveal very little about the space of plausible algorithms in general. This does not seem to be a happy conclusion, for those cognitive scientists who build and study backpropagation trained networks. However, I now want to make a case that things may not be as grim as they may at first appear.

Biology and Bias

In order for the argument to proceed further, it is necessary to bring in another premise. Gould (1980: p. 26) cites Francis Jacob as the source of the aphorism that "Nature is an excellent tinkerer, not a divine artificer". The crucial point here is that the evolutionary process is not one that produces perfect solutions, in some sense, to particular problems. Rather, evolution tends to develop solutions to problems that work, even if the solutions themselves are sub-optimal from a design perspective. Gould (1980) argues this point by citing the pseudo-thumb of the Giant Panda, along with several other examples.

As there is a potential for some confusion at this point, it is worth pausing briefly to consider two ways in which solutions to problems can be evaluated. The first way to think of a solution to a particular problem or set of problems, is from what I term (for want of a better term) an 'design' perspective. Suppose some divine artificer was to wish to provide a Giant Panda with a means of holding bamboo shoots. Such an artificer would presumably be in a position to design a solution which would be as perfect as possible, in terms of simplicity, efficiency, robustness and so on, given the constraints of the problem at hand. Note though that the artificer (being divine) would not be constrained with respect to the materials from which the solution could be fashioned. Perhaps in the instance of the Panda's Thumb, an extra digit would be the best way of solving the problem. Compare this to the actual solution developed to the problem through the evolutionary process. In the case of the Giant Panda, the pseudo-thumb is actually created by the extension of a bone in the wrist. Such a solution respects the fact that there are only certain resources available from which the additional functionality can be derived. However, it may well be the case that such a solution may not be as advantageous as the option of simply adding an extra digit, and consequently may be judged to be 'sub-optimal' when compared to a 'designed' solution.

The relevance of this premise to the current issue is that it suggests something about the kinds of models that are likely to be strongly equivalent. Presumably, what is true of the evolution of parts of the body, is also likely to be true of the mind and brain (C.f.

Cosmides and Tooby, 19**). If biological bodies are quirky and sub-optimal in some respects when considered from a design perspective, then it is not unreasonable to assume that the structures that govern biological cognition are similarly idiosyncratic. However, if this is the case, then it would seem that there is a very real problem that has to be faced by researchers who are attempting to model cognition. The problem is to find a way of generating models with the appropriate kinds of idiosyncrasies (whatever they may be).

In practical terms, dealing with this problem is not too easy. The reason for this is that the standard process of training which researchers go through at the undergraduate and the graduate level is antithetical to idiosyncrasy. When a person takes their first class in programming, one of the first lessons learned is to always try and find 'elegant' solutions to problems. Similarly, in a logic class students are often penalized for deriving proofs that are overly long or clumsy, even if the proofs themselves do not contain any erroneous application of the rules of inference. Analogous examples can be found easily enough in almost any discipline. The point is, through the formal process of education, most researchers are inured to do the exact opposite of what seems to be suggested by the evolutionary record. We are trained to favor well-designed solutions over sub-optimal, kludgy ones. Thus, it is difficult (or at least highly counter-intuitive) to figure out ways of constructing models of cognitive function which are idiosyncratic.

It should be made plain though that the claim here is not that researchers *cannot* produce cognitive models of the appropriate kind. It is just that doing so is not a straightforward or obvious process. There is one exception to this though. If a researcher leaves the selection of key features of a model to some mechanical process, then the tendency to avoid certain types of solutions to problems (i.e. 'messy' solutions) can be overcome. Provided that a model meets the minimum requirement, that it performs in a manner which is at least weakly equivalent to the biological system which it is supposed to emulate, then the optimality of the solution deployed is not initially an issue. The proposal I wish to make here is that this situation may, in fact, end up favoring the kinds of solutions to cognitive problems discovered by connectionist networks trained using the backpropagation learning procedure. This is because when the hidden units of a trained become sensitive to certain input properties whilst learning to solve a problems set, there are no prior constraints upon the selected set of input properties, other than the fact that they must serve to solve the problem at hand.

There is an immediate an obvious objection to this proposal: "Doesn't this end up putting cognitive scientists who train connectionist networks using backpropagation into a position of effectively looking for a proverbial needle in a haystack, when it comes to finding models which can be informative about biological cognition?" The fact noted earlier, that there are potentially a very large number of algorithms for computing a particular function, seems to suggest that this will be the case. Although this objection seems plausible at first, there are *prima facie* reasons to believe that, in practice, it may not actually present as much of a barrier to progress as it initially appears.

The first response to this objection is to note that it is by no means clear that there actually will be a large number of algorithms for computing a particular cognitive function. It may turn out to be the case that there are comparatively few, or even just one. This is ultimately a type of question that needs to be treated empirically. For example, Berkeley *et al*. (1995) trained a network to determine the validity and type of a set of logic problems, originally studied by Bechtel and Abrahamsen (1991). The detailed analysis of this network revealed that the network had developed 'rules' which were in many instances close analogies to the classical rules of natural deduction (Cf. Bergmann, Moor and Nelson 1990). Perhaps the traditional rules of inference are the only way of successfully determining validity. There is evidence that similar cases exist in other problem domains too. For instance, Lehky and Sejnowski (1988, 1990) trained a network to determine three-dimensional shapes from shaded forms. They determined that the hidden units had become tuned to features that were remarkably similar to those found in real neurons in the visual cortex (see Spitzer, 1999: pp 60-62, and Churchland and Sejnowski, 1992: pp. 183-188 for a detailed discussion of these results).

The second response to the objection depends upon understanding the role of hidden units in trained networks as functioning as detectors of input properties which are needed to solve the particular set of problem at hand. If it is determined empirically that all networks which learn to solve a particular set of problems are sensitive to some particular set of input properties, then there may be grounds for hypothesizing that biological cognitive agents are sensitive to the same properties. This is just the kind of hypothesis which could be (at least in principle) verified by conducting studies on biological subjects. The network methodology would act as a means of generating hypotheses about the particular function in question.

The third response to the objection is that there are a number of performance criteria, such as the ability of networks to generalize to new data, which could easily be deployed in order to determine the effectiveness of the solution to a problem found by a network. This too would offer a ready and easy means of determining which algorithms were worthy of further study and which were not. In addition, all researchers, be they interested in connectionist modeling, or modeling in other ways, have a duty to compare the behaviors of their systems with the behaviors of biological systems, before making claims about biological cognition on the basis of their models. Moreover, the behaviors of the system should include more than just the system's behavior on the task explicitly at hand. That is to say, emergent behaviors of the systems should also be considered and assessed. This, after all, is one of the crucial (though regrettably, all too often overlooked) steps in determining whether or not a model is strongly equivalent to biological systems. This equivalence is required in order to justify direct inferences based upon any kind of computational model. These kinds of considerations would assist in determining which models provided good evidence and which did not, thereby limiting the size of the algorithmic space that needed to be investigated.

Of course, there is no guarantee that any of these responses would be helpful in the case of all particular cognitive functions. Whether or not this was the case, is a matter that

would have to be determined empirically. The point that needs to be appreciated here is that the objection is not fatal to the proposed research strategy, until such time as some studies have been done and some evidence collected. Moreover, adopting this methodological strategy provides a viable and adequately justified role for connectionist research using backpropagation .

Conclusion

In the above, I have attempted to sketch a means of justifying connectionist research using the backpropagation learning procedure. The case has briefly been made that we may be able to use models of this kind, when weakly equivalent, to discover salient facts about cognitive systems in general. It has also been suggested that it is possible that connectionist systems may be able to do more than this. However, throughout the argument, there has been one assumption made that needs to be made explicit. This is the assumption that there is some viable means of determining exactly which input properties the hidden layer of processing units becomes sensitive to, when a network has learned to solve a problem. At the current time, this is a controversial and problematic issue (see McCloskey 1991), which cannot be discussed further here. However, assuming that connectionists can use backpropogation networks to recover sets of properties that can solve particular cognitive problems of interest, then it seems that they have some justification for their methodology.

Bibliography

Bechtel, W. and Abrahamsen, A. (1991) Connectionism and the Mind, Basil Blackwell (Cambridge, Mass.).

Bergmann, M., Moor, J., & Nelson, J. (1990), The Logic Book McGraw-Hill (New York).

Berkeley, I., Dawson, M., Medler, D., Schopflocher, D., and Hornsby, L. (1995), "Density Plots of Hidden Unit Activations Reveal Interpretable Bands", in Connection Science, 7, pp. 167-186.

Brassard, G. and Bratley, P. (1988), Algorithmics, Theory & Practice, Prentice Hall (New York).

Churchland, P. M. (1989), The Neurocomputational Perspective: The Nature of Mind and the Structure of Science, MIT Press (Cambridge, Mass.).

Churchland, P. S. and Sejnowski, T. (1992) The Computational Brain, MIT Press (Cambridge, MA).

Clark, A. (1989), Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing, MIT Press (Cambridge, Mass.).

Clark, A. and Thornton, C. (1997), "Trading Spaces: Computation, representation, and the limits of uninformed learning", in Behavioral and Brain Sciences, 20, pp.57-90.

Cosmides, L. and Tooby, J. (1994), “From Function to Structure: The Role of Evolutionary Biology and Computational Theories in Cognitive Neuroscience” in Cosmides, L. & Tooby, J. 1994. Gazzaniga, (1994).

Cummins, R. (1989), Meaning and Mental Representation, MIT Press (Cambridge, Mass.).

Dennett, D. (1978), Brainstorms: Philosophical Essays on Mind and Psychology, Bradford Books (Montgomery, VT).

Dennet, D. (1995), Darwin's Dangerous Idea : Evolution and the Meanings of Life, Simon & Schuster (New York).

Gazzaniga, M. (Ed.), (1994), The Cognitive Neurosciences., MIT Press (Cambridge, MA).

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

Gould, S. (1980), The Panada’s Thumb: More Reflections in Natural History, Norton & Co. (New York).

Haugeland, J. (1985), Artificial Intelligence: The Very Idea, MIT Press (Cambridge, Mass.).

Quinlan, P. (1991), Connectionism and Psychology: A Psychological Perspective on New Connectionist Research, U of Chicago Press (Chicago, IL).

Lehky, S. and Sejnowski, T. (1988), "Network model of shape-from-shading: neural function arises from both receptive and projective fields", in Nature, 333: pp. 452-452.

Lehky, S. and Sejnowski, T. (1990), "Neural network model of visual cortex for determining surface curvature from images of shaded surfaces" Proceedings of the Royal Society of London B, 240: pp.251-278.

McClelland, J. Rumelhart, D. and Hinton, G. (1986), “The Appeal of Parallel Distributed Processing”, in Rumelhart *et al*. (1986: pp. 3-44).

Pylyshyn, Z. (1984), Computation and Cognition, MIT Press (Cambridge MA).

Rumelhart, D., McClelland, J. and The PDP Research Group (1986), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, (2 Vols.), MIT Press (Cambridge, Mass.).

Spitzer, M. (1999), The Mind Within the Net: Models of Learning, Thinking, and Acting, MIT Press (Cambridge, Mass.).

Sterelny, K. (1990), The Representational Theory of Mind, Blackwell (Oxford).