Preface to the Special Issue on Connectionist Symbol Processing

Geoffrey E. Hinton
Department of Computer Science, University of Toronto, 10 Kings College Road, Toronto, Canada M5S 1A4

Connectionist networks are composed of relatively simple, neuron-like processing elements that store all their long-term knowledge in the strengths of the connections between processors. In the last decade there has been considerable progress in developing learning procedures for these networks that allow them to automatically construct their own internal representations [6–8, 10]. The learning procedures are typically applied in networks that map input vectors to output vectors via a few layers of "hidden" units. The network learns to dedicate particular hidden units to particular pieces or aspects of the input vector that are relevant in determining the output. The network generally learns to use distributed representations [5] in which each input vector is represented by activity in many different hidden units, and each hidden unit is involved in representing many different input vectors.

Within the connectionist community, there has been a long and unresolved debate between those who favor localist representations in which each processing element corresponds to a meaningful concept [3, 11] and those who favor distributed representations. The major criticism of distributed representations has been that they cannot handle structured knowledge properly and this criticism has motivated many of the papers in this issue. Another criticism has been the unintelligibility of distributed representations. As soon as there are several hidden layers, it becomes very difficult to say what each hidden unit is representing. Other things being equal, it is clearly desirable to understand *how* a system performing a task such as medical diagnosis arrives at a particular conclusion and to provide this information to the user. A large pattern of activities or set of learned weights is not a convincing explanation. If, however, the large set of weights performs consistently better than an alternative system that can explain its reasoning, it might be better to settle for the system that works best. Under certain conditions, we can be quite justified in trusting a system even if we have very little understanding of how it arrives at a particular conclusion. Using the probably approximately correct framework developed in [12], Baum and Haussler [1] have shown that if a neural network can be

Artificial Intelligence **46** (1990) 1–4

adapted to produce the correct answer for a number of training cases that is large compared with the size of the network, it can be trusted to respond correctly to previously unseen cases provided they are drawn from the same population using the same distribution as the training cases. This remarkable result undermines the common idea that explanations are a *necessary* feature of trustworthy systems.

Unfortunately, the kinds of networks in which the learning procedures have generally been applied lack some properties that AI researchers working within the symbolic paradigm consider to be essential in a general-purpose information processing system [4]. The ability to represent complex hierarchical structures efficiently and to apply structure sensitive operations to these representations seems to be essential. Most connectionist researchers accept this, though they expect that this ability may be implemented in ways that have not been anticipated within the standard symbol-processing tradition. Moreover, they hope that the connectionist approach will be far better at dealing with interactions between levels. Many of the challenging phenomena in language, for example, have to do with cross-over phenomena, in which details at one level have consequences for details at another. Such phenomena are often difficult to capture within the more traditional framework.[1]

Most connectionist researchers are aware of the gulf in representational power between a typical connectionist network and a set of statements in a language such as predicate calculus. They continue to develop the connectionist framework not because they are blind to its current limitations, but because they aim to eventually bridge the gulf by building outwards from a foundation that includes automatic learning procedures and/or massively parallel computation as essential ingredients. Subject to these hard constraints, they aim to progressively improve representational power. The papers in this special issue should be interpreted from that perspective. It is not the standard AI perspective in which the ability to succinctly represent and efficiently apply complex knowledge is viewed as a more important consideration than automatic learning.

There have been important battles in the past between symbolic AI researchers who focussed on representational power and other researchers who nailed their flag to automatic learning procedures. The perceptron battle was a resounding victory for symbolic AI. A single layer of adaptive linear threshold units was just too limited, and no effective learning procedure was then known for multilayer networks. The subsequent speech recognition battle between symbolic AI and those who believed in adaptive hidden Markov models (HMMs) is not as commonly mentioned in AI circles. It turned out that the complex, hand-designed representations and rules in systems like HEARSAY [9] were no match for HMMs even though HMMs, being a variety of finite

[1] Elman, Personal communication.

state machine, are clearly very limited in representational power. The outcomes of these two battles suggest that as the learning procedures become more sophisticated the advantage of automatic parameter tuning may more than outweigh the representational inadequacies of the restricted systems that admit such optimization techniques. An optimal member of a class of incorrect models may work much better than a far from optimal member of a class that contains the right model. Clearly, the ultimate goal is efficient learning procedures for representationally powerful systems. The disagreement is about which of these two objectives should be sacrificed in the short term.

Current connectionist learning procedures such as backpropagation are comparable in power to the learning procedure for HMMs. Indeed, one kind of backpropagation network is equivalent to one kind of hidden Markov recognizer [2]. As further theoretical progress is made, we can expect the optimization techniques used for connectionist learning to become much more efficient and, if these techniques can be applied in networks with greater representational abilities, we may see artificial neural networks that can do much more than just classify patterns. But for now, the problem is to devise effective ways of representing complex structures in connectionist networks without sacrificing the ability to learn the representations. My own view is that connectionists are still a very long way from solving this problem, but the papers in this issue suggest some interesting directions to pursue.

REFERENCES

1. E.B. Baum and D. Haussler, What size net gives valid generalization? *Neural Comput.* **1** (1989) 151–160.
2. J.S. Bridle, Alpha-nets: A recurrent "neural" network architecture with a hidden Markov model interpretation, Tech. Rept. SP Research Note 104, Royal Signals and Radar Establishment, UK (1989); also *Speech Communication* (to appear) Special *Neurospeech* Issue.
3. J.A. Feldman, Neural representation of conceptual knowledge, Tech. Rept. TR189, Department of Computer Science, University of Rochester, Rochester, NY (1986).
4. J.A. Fodor and Z.W. Pylyshyn, Connectionism and cognitive architecture: A critical analysis, *Cognition* **28** (1988) 3–71.
5. G.E. Hinton, J.L. McClelland and D.E. Rumelhart, Distributed representations, in: D.E. Rumelhart, J.L. McClelland and the PDP Research Group, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* **1**: *Foundations* (MIT Press/Bradford Books, Cambridge, MA, 1986) 77–109.
6. G.E. Hinton and T.J. Sejnowski, Learning and relearning in Boltzmann machines, in: D.E. Rumelhart, J.L. McClelland and the PDP Research Group, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* **1**: *Foundations* (MIT Press/Bradford Books, Cambridge, MA, 1986) 282–317.
7. T. Kohonen, Self-organized formation of topologically correct feature maps, *Biol. Cybern.* **43** (1982) 59–69.
8. J. Moody and C. Darken, Fast learning in networks of locally-tuned processing units, *Neural Comput.* **1** (1989) 281–294.
9. D.R. Reddy, L.D. Erman, R.D. Fennell and R.B. Neely, The hearsay speech understanding system: An example of the recognition process, in: *Proceedings IJCAI-73*, Stanford, CA (1973) 185–194.

10. D.E. Rumelhart, G.E. Hinton and R.J. Williams, Learning internal representations by back-propagating errors, *Nature* **323** (1986) 533–536.
11. L. Shastri, A connectionist approach to knowledge representation and limited interference, *Cognitive Sci.* **12** (1988) 331–392.
12. L.G. Valiant, A theory of the learnable, *Commun. ACM* **27** (1984) 1134–1142.