Machine Learning 7, 105 - 107 (1991)
© 1991 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

Introduction

The goal of this special issue is twofold: first, to acquaint members of the machine learning community with the latest results in connectionist language learning, and second, to make these five inter-related papers available in a single publication as a resource for others working in the area. In the remainder of this introduction I will sketch what it is that I think the connectionist approach offers us, and how the papers in this special issue advance the state of the art. But this is not going to be a cheerleading piece about the wonders of "brain-style computation" and the imminent death of symbolic AI. Rather, I hope to tempt the reader into examining some novel ideas that expand the current scope of AI.

"Connectionism" is not a single distinct approach to learning. It is a collection of loosely-related ideas covering a broad range of intellectual territory: dynamical systems theory, computational neuroscience, cognitive modeling, and so on. While the ultimate goal of the connectionist enterprise is supposed to be discovering how brains embody intelligence, current connectionist learning models are nowhere near achieving either general intelligence or neural plausibility. (AI as a whole is still far from producing generally intelligent agents, so this is not an indictment of any one particular approach.)

Despite the fact that today's "neural net" models are not brain-like to any meaningful extent, they have interesting insights to offer the Machine Learning community. The connectionist approach to computation is quite different from discrete symbol manipulation. In some cases, such as the paper by Mozer and Bachrach, this difference manifests itself in useful behavioral properties and broader/alternative conceptualizations of the learning problem. In other cases, such as the paper by Servan-Schreiber, Cleeremans, and McClelland, and the following one by Elman, it has led to some surprising suggestions about human cognition. The papers in this special issue represent some of the best connectionist work to date on the problem of language learning. They span the range from solid theoretical analysis (Porat and Feldman) to bold conjecture (Pollack). Together they demonstrate the richness of this exciting multi-disciplinary movement.

So what exactly is the connectionist approach to computation? Common themes have been massive parallelism and simple computing units with limited communication. Beyond these general properties, connectionist models can be divided into two major classes. In localist models, individual units represent specific items or concepts. Localist models are generally constructed by hand, or in the case of Porat and Feldman's work, by a discrete learning algorithm. The principal advantages of localist models are that they are easy both to construct and to analyze.

Distributed models represent information as diffuse patterns of activity of a collection of units, and are usually constructed by gradient-descent learning algorithms such as back-propagation. The principal advantages of these models are that they develop their own representations *de novo*, and naturally generalize to novel inputs. A disadvantage is that their behavior is harder to control; typically many iterations of the learning algorithm are required to achieve acceptable performance on the training set, and generalization ability must be tested empirically. A special case of distributed models, the Simple Recurrent Network

(SRN), can process sequences of symbols by inputting them one at a time. Recurrent or "feedback" connections allow the SRN to incrementally construct representations for sequences, or whichever bits of them are relevant, as locations in a high-dimensional vector space.

In the first paper in this special issue, Porat and Feldman consider the tractability of learning a regular language (or equivalently, a finite state automaton) when input examples are lexicographically ordered. Although the construction of a minimum-state deterministic FSA is known to be NP-hard, they show that the problem can be solved in polynomial time, with limited memory (no storage of previously-observed training examples), when lexicographic ordering is assumed. At the conclusion of the paper they describe an implementation of their algorithm by a localist-type connectionist network.

Mozer and Bachrach are also concerned with the efficient induction of finite state automata. They consider the problem of a robot wandering around in a simulated finite-state environment that must learn to predict the effects of its actions. A symbolic learning algorithm by Schapire and Rivest solves this problem using a structure called an *update graph*. Mozer and Bachrach show that a connectionist implementation of the update graph idea, based on backpropagation learning, learns small worlds more quickly than the original algorithm. Furthermore, the discrete symbolic update graph structure turns out to be merely a limiting case of a more general, continuous representation created by backprop through gradient-descent learning.

Approximation of discrete, finite-state phenomena by continuous dynamical systems is a topic of much current research. The next three papers in this special issue analyze the internal representations created by various recurrent networks. Servan-Schreiber, Cleeremans, and McClelland train an Elman-style Simple Recurrent Network on strings from a moderately complex regular language. They use hierarchical clustering to analyze the hidden unit activation patterns, revealing a complex state structure that is richer than the minimal-state DFSA describing the same data. Perhaps their most striking result is that, for certain problems involving one FSA embedded in two places inside another, the resulting grammar is not learnable due to the difficulty of keeping the two embeddings separate. However, if the arc transition probabilities are altered from the usual .5/.5 to, say, .6/.4 for one embedding and .4/.6 for the other, this statistical marker distinguishes the two sufficiently so that the grammar can be learned. The suggestion Servan-Schreiber, Cleeremans, and McClelland make is that statistical variation may also be an important cue in human language learning. For example, although English permits arbitrary embedded clauses, in normal conversation the distribution of syntactic structures for embedded clauses is different from the distribution of structures for surface clauses. This might be a source of unconscious cues to help language learners master embedding.

Elman's paper provides additional insight into embedding phenomena. He considers the problem of learning context-free rather than regular grammars from examples. Recognizing context-free languages requires some sort of stack. The Simple Recurrent Network can learn to simulate a limited-depth stack in the structure of its hidden unit states. What's particularly interesting is that the automaton constructed by the SRN generalizes to sentences that involve slightly more complex structures than the ones in the training set. One cannot hope to get this sort of generalization from a purely symbolic FSA induction algorithm, since it results from the existence of additional states, not required by the training set, that arise automatically as a result of the recursive structure of the task and the continuous nature of the SRN's state space. Elman also introduces a new graphical technique for studying network behavior based on principal components analysis. He shows that sentences with multiple levels of embedding produce state space trajectories with an intriguing self-similar structure.

The development and shape of a recurrent network's state space is the subject of Pollack's paper, the most provocative in this collection. Pollack looks more closely at a connectionist network as a continuous dynamical system. He describes a new type of machine learning phenomenon: induction by phase transition. He then shows that under certain conditions, the state space created by these machines can have a fractal or chaotic structure, with a potentially infinite number of states. This is graphically illustrated using a higher-order recurrent network trained to recognize various regular languages over binary strings. Finally, Pollack suggests that it might be possible to exploit the fractal dynamics of these systems to achieve a generative capacity beyond that of finite-state machines.

It remains to be seen whether dynamical systems theory will permanently alter our understanding of symbolic phenomena, or merely provide an interesting diversion from classical automata theory. Our understanding of symbol processing in recurrent connectionist networks is still at an early, almost pretheoretic stage. Conjectures about how an infinite state space relates to Turing machines—or to real neural representations—are admittedly premature at this point, but they are tantalizing nonetheless. By studying these papers, the reader can share in the excitement of working at the connectionist frontier.

David S. Touretzky
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890