To appear in: *Artificial Intelligence*

Book Review

David Touretzky, Jeffrey Elman, Terrence Sejnowski, and Geoffrey Hinton, eds.

Connectionist Models: Proceedings of the 1990 Summer School[1]

Reviewed by: **Subutai Ahmad**
Siemens Central Research, ZFE ST SN61,
Otto-Hahn-Ring 6, 8000 Munich 83, Germany.
Electronic mail: ahmad@icsi.berkeley.edu

(Received February 1992; Revised July 1992)

New Kids on the Block

Connectionist Models is a collection of forty papers representing a wide variety of research topics in connectionism. The book is distinguished by a single feature: the papers are almost exclusively contributions of graduate students active in the field. The students were selected by a rigorous review process and participated in a two week long summer school devoted to connectionism[2]. As the ambitious editors state in the foreword:

> The forty papers in this volume exemplify the tremendous breadth of research under way in the field of connectionist modeling. ... The papers selected for this proceedings offer an intense, pithy snapshot of the state of the art in 1990. ... [The students] are among the best and the brightest in the neural nets game. (Page vii)

These are bold claims and, if true, the reader is presented with an exciting opportunity to sample the frontiers of connectionism. Their words imply two ways to approach the book. The book must be read not just as a random collection of scientific papers, but also as a challenge to evaluate a controversial field.

[1] (Morgan Kaufmann, San Mateo, CA, 1991); 404 pages, US$?? (paperback). ISBN 0-55860-156-2.

[2] This summer school is actually the third in a series, previous ones being held in 1986 and 1988. The proceedings of the 1988 summer school (which I had the priviledge of participating in) are reviewed by Nigel Goddard in [4]. Continuing the pattern, a fourth school is scheduled to be held in 1993 in Boulder, CO.

The editors are right to be proud of this volume. The quality of the papers is high, being generally well-written, concise, and insightful. The book, however, is not an easy collection to digest. There is a wide range of topics with the papers varying significantly in the level of sophistication required of the reader. The book is organized into ten parts, with headings that range from "Mean Field, Boltzmann, and Hopfield networks" to simply "Biology". Within these sections there are articles which require previous experience with a particular methodology. Consider Dayan's article on reinforcement learning which explores a subtle second order effect in a learning algorithm proposed by Sutton. The conscientious reader will discover an interesting analysis of an important learning scenario but this does require some work. (Dayan shows that there is a second order effect but perhaps not a significant amount.) On the other hand there are several articles written for audiences with more diverse backgrounds. In particular, the articles by Miller and Todd, Goodhill, Wieland, Plunkett *et al*, and Lange *et al* provide good introductions to their topics before delving into details.

It would be foolhardy to attempt a review of every paper in the collection. Instead, the review will focus on sets of papers which contribute to some important themes in a coherent fashion. These themes have played important roles in non-connectionist AI as well as connectionism and as such provide good benchmarks for evaluating the field.

Connectionist Language and Cognitive Processing

The largest section of the book is devoted to connectionist models of language and cognitive processing. A central theme concerns the claim that many cognitive faculties thought to be dominated by rule-based systems can also be explained as a structured system of *mappings* from input to output vectors. We'll consider three such papers and one that represents a different viewpoint.

Plunkett *et al* have written a very good article on the learning of verb morphology in children. Several experiments have shown that children go through three distinct phases in learning the past tenses of verbs. In the first phase children produce past tenses correctly, even for irregular verbs. During the second phase, they start making mistakes, typically overgeneralizing rules for regular verbs to irregular verbs (e.g. "I goed home."). After some time there is a transition to the final phase, where there are no more errors. The traditional explanation for this is that during the first phase children learn by rote, storing past tenses in a lexicon. After some time they start incorporating rules governing the production of past tenses, but haven't yet learned the exceptions. Finally they learn both the exceptions and the rules. Rumelhart and McClelland[8] showed that such behavior can be reproduced by a feed-forward network learning mappings between verbs and their past tenses. However their model has been beset by controversy, particularly due to their unnatural training scheme. The Plunkett *et al* paper presents a much more thorough set of experiments that consider the effects of some of these factors. The key result is that the learning curves of children can be reproduced under a large variety of training regimes. They clearly show that such behavior can emerge from a general purpose learning algorithm extracting regularities from the environment. They make the point that a single system is capable of producing the desired behavior (as opposed to a dual system consisting of a lexicon plus a rule-extracting mechanism). Their final conclusion however seems a bit puzzling:

> Thus, evaluations of network performance on novel verbs (in comparison to patterns of errors across learning) suggests that the network, like the child, is best characterized as a rule-governed system. (Page 217)

The whole point of the Rumelhart and McClelland paper, which they clearly support, is to prove the opposite: that the acquisition of verb morphology, a seemingly rule-governed system, can be characterized as a general network learning a set of mappings from vectors to vectors.

This general theme is developed further by Touretzky, one of the faculty contributers. His paper models part of the process of deriving surface phonetics from underlying phonemes. He shows that a highly constrained set of feed-forward networks of limited depth can be used to implement these derivations. The twist here is the emphasis on the word "constrained". Unlike the previous paper, Touretzky's network is not a general purpose one. It can perform only a limited set of mappings. Their hypothesis is that this set is exactly equal to those manipulations seen in human languages. The model is one where predictions can be stated precisely and therefore one that can be easily refuted. It will be interesting to see how well it holds up in the future.

Jennings and Keele present a model of sequence learning based on simple recurrent networks. They show that by appropriately picking input patterns one can reproduce the response curves of people in a variety of sequence learning tasks. The paper models experimental results which show that certain sequences cannot be learned by humans without some form of attention. In particular sequences where the next item is not uniquely predicted by the previous item are difficult to learn while performing a distracting task (e.g. 1-3-2-3-1-2 is difficult but 1-3-2-1-3-2 is easy). A recurrent network with extra "plan inputs" is trained with such sequences. Without additional information, the network exhibits difficulties in learning sequences of the first type. When the plan units encode extra hierarchical information about the current block within a sequence, the network learns faster. The claim is that this corresponds to an attention-based hierarchical parsing of sequences by humans. Unfortunately this article demonstrates some of the dangers of blindly applying connectionist networks to psychophysics. For example, there is no convincing argument that hierarchical inputs are required. There is probably a variety of ways to get the network to learn the difficult sequences. The network might exhibit the same behavior if *random* vectors are associated with each block. In addition, the learning curves are likely to fluctuate quite a bit depending on the number of hidden units, the learning parameters, and so on. A more careful analysis of such factors (none of the above are mentioned in the paper) would greatly strengthen the claims. As it stands, the relationship between the network's behavior and human behavior is not at all clear.

In contrast to the previous three papers, the article by Lange *et al* describes a highly structured frame-based connectionist system. The flavor is more akin to semantic networks than vector mappings. The paper presents a well thought out integrated model of language understanding and analogical inferencing (structural and semantic). The main feature here is the ability to perform relatively complex linguistic inferences in parallel. The authors construct a network whose connectivity structure reflects the underlying predicates and the relationships between them. The key to this process is a scheme for dynamic variable binding using unique patterns of activity. With this tool, sets of variable instantiations are propagated through the network. When the network settles, nodes with the highest confidence values represent the most plausible interpretation. Although it is difficult to imagine the network scaling up to deal with whole stories, the authors do present impressive examples using short sequences of ambiguous sentences. Unfortunately the description of the complete network is somewhat involved and the few pages allocated to it are hardly sufficient. To get a complete understanding, the reader should look up some of the references to their other work.

Learning and Generalization

Almost all of the papers in this book incorporate learning to some extent. Included among these are a good set of papers that study learning itself. Here we consider two that deal with the following question: given a fixed training set, what can we say or do about the eventual generalization?

Back-propagation is the learning algorithm used in most connectionist models. When there is a large set of training examples, the learning algorithm usually leads to a network that generalizes very well on future unseen examples. The analysis by Hampshire and Pearlmutter suggest why this is the case. They show that feed-forward networks, when trained as a classifier, act as optimal Bayesian discriminant functions under a variety of situations. For example back-propagation using the sum-squared error measure leads to networks whose outputs encode the maximal *a posteriori* probabilities of the classes. Unfortunately it is not clear yet how this applies to practical situations: their analysis is currently limited to the cases where training data tends to infinity and the training examples are statistically independent.

The second article concerns itself with the opposite case: when the training set is small. In such situations, there are often several local minima in the error surface but only some of these correspond to good solutions. During the 1988 summer school, Rumelhart suggested an intriguing modification of back-propagation designed to attack this problem. The paper by Weigend, Rumelhart and Huberman in this collection shows that, at least in one domain, the method works very well and can outperform standard statistical techniques. The basic idea is to use a version of Occam's principle: when faced with a choice of networks, prefer smaller networks over larger ones. The trick is to augment the standard error measure with a term that penalizes networks with large weights. The resulting learning rule performs gradient descent in a measure of network complexity as well as training error. One of the interesting features is that it tends to automatically eliminate unnecessary parameters and hidden units. The technique is applied to the prediction of sunspot data, a benchmark problem in statistics. Surprisingly, the method outperformed the best current statistical solution. To this day, this remains one of the few results where a connectionist network does substantially better than the best known traditional solution to a problem.[3]

Unsupervised Learning and Clustering

A longstanding and fascinating theme in connectionist learning has been the automatic development of features through the use of unsupervised learning rules[4]. These proceedings contain several good articles which explore various aspects of this topic. They are a bit scattered around in the book but when read together provide a coherent picture of many of the issues.

Several researchers (e.g. Linsker[5]) have shown that variants of the general Hebb rule can lead to the development of feature detectors similar to those found in mammalian visual cortex. One of the basic questions is: Is there a single learning rule that can account for all cortical development? One of the best written papers in this collection, by Goodhill, explores this issue. The paper considers two somewhat distinct processes: the formation of topographic maps (a common biological phenomenon where neurons physically near one another respond only to visual stimuli that are similarly near each other) and ocular dominance columns (where bands of neurons respond exclusively to stimuli from a single eye). These are two areas with long histories of fruitful interaction between modeling and experimentation. The paper contains a very good review of some of this work, both biological and computational. Goodhill then presents a theoretical framework for dealing with both processes simultaneously. Simulation results show that the model accounts for effects seen under both normal and abnormal development. Although it cannot account for all the experiments, perhaps we are one step closer towards a grand unified learning rule.

[3] See the paper by Nowlan (a participant of the 1988 summer school) and Hinton[6] for a related approach to this problem. For another famous example of a connectionist network improving on conventional solutions see the recent paper by Tesauro[10] (a participant of the 1986 summer school) describing a neural network tournament-level backgammon program.

[4] The interested reader should consult the excellent review article by Becker[2], a 1988 summer school graduate.

Why should there be a single learning rule? One answer is that some learning rules lead to optimal feature detectors so one rule might suffice. Oja[7] has shown that a linear Hebb neuron adapts to extract the first principal component of the covariance of its inputs. In other words, such a neuron learns to detect the feature which lies along the axis of maximal variance in input space. The article by Levin extends upon this work. He shows how a set of such neurons can be networked to extract the first n principal components. Levin's analysis proves that his scheme will converge with high probability, but it is currently limited to linear neurons. Cotrell's paper briefly discusses a non-linear, supervised scheme for extracting principal components. As Linsker has argued, such techniques lead to feature detectors which maximally preserve input information.

But the issue is far from resolved. Any discussion of optimality is meaningless without defining the word "optimal", and this is likely to be goal-dependent. Intrator argues convincingly that preferring the axis with maximal variance is often the wrong thing to do if the ultimate goal is classification. He emphasizes the difference between features used for classification (where we want to distinguish between classes) and features used for function estimation or data compression. Consider a set of points which are widely distributed over the y-axis but are restricted to two narrow intervals close together on the x-axis. A feature using the x-axis is clearly best able to separate the cluster. Picking the axis with maximal variance however would choose the y-axis. Instead Intrator proposes a complementary learning rule, derived from a cost function that prefers multi-modal distributions. To demonstrate its usefulness, he shows that the features found by this learning rule perform better on a speech phoneme classification task. It is unclear though, how different this new learning rule really is. For example, it is easy to concoct an input distribution such that the two classes of rules generate identical sets of features (i.e. if the axes with large variance also contained the most multi-modal distributions). It would certainly be worth studying the features that evolve if such a network is given random inputs, as in Linsker's work. Interestingly, Intrator's learning rule turns out to be a modification of a rule that has also been used to explain many aspects of cortical development, including changes in ocular dominance columns[3].

Planning and Reinforcement Learning

When it comes to generating complex compositional plans, connectionist methods have yet to reach a high level of performance. There is no system close to the sophistication of even early AI planners such as STRIPS. Part of the reason is that the emphasis has been on the difficult problem of *learning* the correct sequence of actions necessary to reach a given goal. The key problem to be solved here is the temporal credit assignment problem: given that the system gets no reinforcement until the end of a (potentially long) sequence, how should one credit the intermediate actions? The following two articles provide good introductions to many of the issues involved.

Algorithms for dealing with temporal credit assignment generally fall into two categories: those which build internal models of the environment, and those which do not. The first type typically learns to predict the effect of actions on the environment and then, given the current state, uses this model to choose the best next action. The second type simply learns a mapping between current state and next action. The article by Barto and Singh directly compares these two methods on a simple task. It is clear the latter method should be computationally faster. The surprising result is that (at least in one simple task) it can also perform better.

A second interesting article on this topic is by Schmidhuber, who has tackled a variety of problems dealing with reinforcement learning. Although not terribly well-written and at a somewhat general level, the article provides a good introduction to the works of this (very prolific) author. (It is not often that one sees references of the form Schmidhuber (1990a) through Schmidhuber (1990h)!) Right at the end one finds perhaps the most interesting part of the paper: a proposal for learning sub-goal generation. The task is to learn to generate the correct sequence of sub-goals assuming the sub-goals themselves have already been learned. The idea is to use a separate evaluator network E which has already learned the relationship between sub-goals. Given two sub-goals this network outputs a 1 if the first is reachable from the second. Given start and goal states S and G, the sub-goal generator should only generate sub-goals S_i such that $E(S, S_i)$ and $E(S_i, G)$ equals 1. If not, the errors are used to train the sub-goal generator. Although intriguing, the proposal should be backed up by simulations. Unfortunately we are only given one paragraph which states simply that some simulations were tried and they worked. The procedure as given will also only work for plans with one sub-goal. The author claims that recursive application of it should be capable of generating more complex plans but no details are given. Interested readers are referred instead to (Schmidhuber, 1990h).

Genetic Algorithms

Three of the papers consider applications of genetic algorithms to connectionist networks. By simply representing properties of networks as bit strings any standard genetic algorithm can be invoked. The paper by Wieland shows how genetic algorithms can be used to evolve recurrent neural network controllers for various versions of the pole-balacing problem. The simulations show that good controllers can evolve using only genetic algorithms (no gradient descent search). The paper is useful for two other reasons: it contains a good set of references, and also contains the full set of equations necessary to simulate the standard pole-balancing problem, multiple unjointed poles, and a single jointed pole.

The other two papers discuss how learning itself might evolve through genetic adaptation. The bulk of the paper by Miller and Todd reviews the study of learning and evolution in Psychology. They then make the point that as long as adaptive pressures favor a learning system then one should evolve. They present a rather simplistic simulation proving their point. A more interesting simulation along the same lines is found in Chalmers' paper. In this work bit strings denote coefficients of possible learning rules chosen from a restricted class of linear functions. An "environment" consisting of a number of supervised boolean learning tasks is used to determine the fitness of the learning rules. This is clearly not a natural environment but the result is appealing. The rule that evolves to have the highest fitness turns out to be the well known delta rule (a specialization of back-propagation for one-layer networks). One can argue that success in natural environments is the ultimate test of learning rule optimality. It would be worth trying out a version of this task using unsupervised learning rules. Perhaps something like the Hebb rule will emerge.

Concluding Remarks

There is so much variety here that it is easy to get immersed in any one of the above topics. What do the proceedings teach us about the state of connectionism as a field? One of the best ways to evaluate this is to contrast these papers with older ones. As a participant of the 1988 summer school[11], I am in a position to compare this school with the previous school. I see at least two major differences and at least one similarity with 1988. The 1988 proceedings contained papers that were clearly more experimental in nature whereas these papers contain more theory to back up claims. In particular, the notion of *optimality* is much more prominent here. It appears in various forms, whether it is Bayesian optimality, optimal credit assignment, optimal use of computational resources, optimal learning rules, or optimal feature extraction. A second key difference is that the papers in the 1988 proceedings heavily emphasized feed-forward back-propagation style networks. This time around there is significantly more variation both in the networks and the algorithms used. These are definitely healthy trends and point to a field that is rapidly maturing.

One major similarity to the previous proceedings is that there is still a heavy emphasis on learning, whether it is learning plans, language, sequences, or features. Learning is clearly one of the strengths of connectionism but there are certainly many other aspects of intelligent behavior which could profit from a massively parallel approach. This pre-occupation with learning tends to deemphasize issues such as control structures, search procedures, and knowledge based methods which in turn leads to difficulties with topics like reasoning, plan generation, language understanding, and several complex perceptual tasks. Touretzky's phonological model and Lange *et al*'s model of analogical inference show that non-learning connectionist systems can indeed be interesting[5], and provide hope that connectionist methods can provide insights into these hard

[5] For other examples of such systems see [9, 1].

problems as well.

Despite the above objection it is easy to recommend the book. The atmosphere of intensity and excitement present at these summer schools is clearly reflected in the proceedings. In part due to the efforts of the faculty, many of us from the previous schools are still active in the field. I am sure the same will be true of the 1990 participants. Both the students and the field have benefited from these summer schools. I look forward to the results of the 1993 school and hope it will not be the last one.

References

[1] S. Ahmad and S. Omohundro. Efficient visual search: A connectionist solution. In *Proceedings of the 13th Annual Conference of the Cognitive Science Society.* Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1991.

[2] S. Becker. Unsupervised learning procedures for neural networks. *International Journal of Neural Systems*, 12, 1991.

[3] E.L. Bienenstock, L.N. Cooper, and P.W. Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2:32–48, 1982.

[4] N. Goddard. Book review of Proceedings of the 1988 Connectionist Models Summer School. *Artificial Intelligence*, 53(2):345–353, 1992.

[5] R. Linsker. How to generate ordered maps by maximizing the mutual information between input and ouput signals. *Neural Computation*, 1:402–411, 1989.

[6] S.J. Nowlan and G.E. Hinton. Simplifying neural networks by soft weight-sharing. *Neural Computation*, 4(4), 1992.

[7] E. Oja. A simplified neuron as a principal component analyzer. *Journal of Mathematical Biology*, 15:267–273, 1982.

[8] D.E. Rumelhart and J.L. McClelland. On learning the past tense of English verbs. In J.L. McClelland, D.E. Rumelhart, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the microstructure of cognition,*, volume 2, pages 216–271. MIT Press, 1986.

[9] L. Shastri. A connectionist approach to knowledge representation and limited inference. *Cognitive Science*, 12(3):331–392, 1988.

[10] G. Tesauro. Practical issues in temporal difference learning. *Machine Learning*, 8:257–277, 1992.

[11] D.S. Touretzky, G.E. Hinton, and T.J. Sejnowski, editors. *Proceedings of the 1988 Connectionist Models Summer School.* Morgan Kaufmann, San Mateo, CA, 1989.