separating boundaries between classes gives error rates on optical character recognition lower than neural nets (Boser, Guyon and Vapnik, 1992).

Often the analogies and language used in the NN community obscure the data analytic reality. There is a lack of reflective introspection into how their methods work, and under what data circumstances. But these lapses are more than offset by the complexity, interest, size and importance of the problems they are tackling; by the sheer creativity and excitement in their research; and by their openness to anything that works.

Comment: Neural Networks and Cognitive Science: Motivations and Applications

James L. McClelland

Artificial neural networks have come and gone and come again—and there are several good reasons to think that this time they will be around for quite a while. Cheng and Titterington have done an excellent job describing that nature of neural network models and their relations to statistical methods, and they have overviewed several applications. They have also suggested why neuroscientists interested in modeling the human brain are interested in such models. In this note, I will point out some additional motivations for the investigation of neural networks. These are motivations arising from the effort to capture key aspects of human cognition and learning that have thus far eluded cognitive science.

A central goal of congnitive science is to understand the full range of human cognitive function. During the 1960s and 1970s, when symbolic approaches to human cognition dominated the field, great progress was made in characterizing mental representations and in capturing the sequential thought processes needed, for example, to solve arithmetic problems, to carry out deductive reasoning tasks, even to prove theorems of logic from given axioms. Indeed, by 1980 a general computer program for solving integro-differential equations had been written. These accomplishments are certainly very valuable, yet they still leave many scholars of cognition with the very strong feeling that something very important is missing. Efforts in machine recognition of spoken and visual input, machine understanding of language, machine comprehension and analysis of text, not to mention machine implementation of creative or insightful thought, all continue to fall short. A huge gap remains between the capabilities of human and machine intelligence.

The interest in the use of neural networks among cognitive scientists springs largely from the hope that they will help us overcome these limitations. Although it is true that there is much to be done before this hope can be fully realized, there are nevertheless good reasons for thinking that artificial neural networks, or at least computationally explicit models that capture key properties of such networks, will play an important role in the effort to capture some of the aspects of human cognitive function that have eluded symbolic approaches. In what follows I mention two reasons for this view.

The first reason arises in the context of a broad class of topics that can be grouped under the rubric of "interpretation." A problem of interpretation arise whenever an input is presented to the senses, be it a printed digit, a footprint, a scientific argument or a work of creative expression such as a poem or a painting. The problem is to determine what the thing is or what it is intended to signify. The problem is difficult because the direct data is generally insufficient so that the ability to determine the correct interpretation depends on context.

Let us consider two examples. The first, shown in Figure 1, is from Massaro (1975) and illustrates the role of context in letter recognition. The same input gives rise to two very different interpretations depending on the context in which it occurs. The second comes from very simple stories of a kind studied by Rumelhart (1977):

> Margie was playing in front of her house when she heard the bell on the ice

James L. McClelland is Professor of Psychology and Professor of Computer Science, Department of Psychology, Carnegie Mellon University, Baker Hall 345-F, Pittsburgh, Pennsylvania 15213.

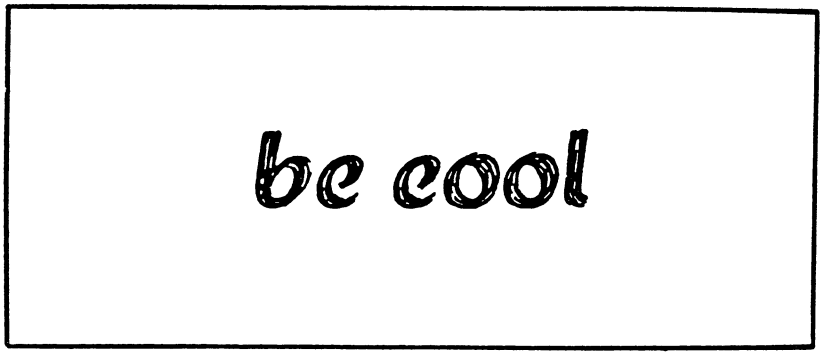

FIG. 1. *The same visual configuration can be interpreted as two different letters, depending on the context. Reprinted with permission from Massaro* (1975) *p.* 382.

> cream truck. She remembered her birthday money and ran inside the house.

In this case, human readers have no trouble figuring out that Margie's birthday money is probably in the house and that she probably ran in to get it so that she could buy herself ice cream. Obviously, this interpretation, engendered by the second sentence of the above story, would not arise if the context were changed:

> Margie lived in a dangerous neighborhood with lots of drug addicts always on the lookout for innocent passers-by to rob. She was coming home from a birthday visit to her grandmother when she saw a couple of the addicts loitering at the corner near her house. She remembered her birthday money and ran inside the house.

What the Massaro and Rumelhart examples have in common is the fact that the direct information—the shape of the character, the words in a sentence—is often not enough by itself to get the correct interpretation. But context is not in general enough by itself—indeed the context often provides only very general and indirect constraints. What one is left with is the sense that it is the aggregated influence of the sum total of the cues rather than any one operating individually that is of crucial importance. Indeed, in real situations it is often the case that ambiguity remains once all the factors have been taken into account. Many psychologists have long argued that it is reasonable to view all acts of interpretation as closely related to Bayesian inference, in that they involve the weighted combination of various direct and contextual cues together with prior biases. Signal detection theory (Green and Swets, 1966), based on a Bayesian analysis of decision making under uncertainty, is a centerpiece of this line of thinking.

As Cheng and Titterington point out, neural networks provide a natural domain for capturing perception and interpretation as probability optimization problems in which direct and contextual information is combined to reach the most likely interpretation given the available input. The use of graded (real-valued) connection weights allows the appropriate weighting of different sources of evidence. The process of settling to a stable attractor state captures nicely the multifaceted nature of most interpretation problems in which the interpretation of one part of an input both influences and is influenced by the interpretation of every other part. Human subjects often behave in ways that are highly consistent with optimal statistical methods (Massaro, 1989) and, indeed, connectionist models that share these properties have been highly successful in accounting for psychological data from perceptual decision tasks (McClelland and Rumelhart, 1981; McClelland and Elman, 1986; McClelland, 1991). A wide range of authors have argued for the use of similar models in sentence comprehension, story understanding, visual scene interpretation and many other related tasks based on the general fact that correct interpretation is not in general possible. The only way to maximize the probability of making the correct decision is to exploit all sources of information.

A second reason why neural networks are relevant to cognitive science arises in the area of learning. Psychological research on learning has gone through many different phases, including a phase lasting from around 1920 to nearly 1960 where it was dominated by stimulus-response theories (in which probabilistic formulations have proven very useful) and another that arose in the 1950s and persisted into the 1960s in which learning was conceptualized in terms of the formulation and testing of deterministic rules, within the symbolic tradition. This approach largely gave way in the 1970s and 1980s to a new approach based on the probabilistic use of accumulated knowledge from examples.

One of the most successful models in this tradition is a model of category learning due to Medin and Schaffer (1978). These authors argued that category learning occurs through the exhaustive storage of all examples in memory. When a test item is presented for categorization, it is compared to all of the examples in memory and each votes for its own category in proportion to its similarity to the test item. The probability of choosing a particular category is equal to the sum of the votes of all of the known exemplars in the category divided by the sum of all of the votes. The key point is that the responses subjects make are probabilistic, not deterministic; and they reflect the influence of specific examples rather than general rules. Neural network models are highly relevant to capturing this kind of learning since each experience leaves its own residue in the form of changes to the connection weights among the units in the network. Indeed, the Medin and Schaffer model can easily be formulated as a neural network model, and a recent, highly successful connectionist model of category learning due to Kruschke (1992) takes just this approach. Kruschke's model makes use of individual units to represent each exemplar and extends the Medin and Schaffer model by using an error correcting learning rule to modify the strengths of the contributions each exemplar makes to the activation of each of the possible categorization responses.

A related difficulty for deterministic rule systems arises in various domains of language. In general, language production and interpretation can both be thought of as mapping problems in which a message in one form of representation must be translated into another form of representation. As two examples, the problem of producing a verb to describe a state or action one wishes to convey, and the problem of producing a spoken sound that corresponds to a written word can both be thought of as mapping problems. In general, in natural languages such problems often involve what might be called quasi-regular—or even better probabilistic—structure. In mapping from spelling to sound, for example, there are important regularities; but at the same time there are many exceptions as well. Often, the exceptions are not simply isolated individual cases but are grouped together in clusters; for example, in English spelling there are many words that violate the rule that EA corresponds to the long E sound as in HEAT; most of these words—THREAD, TREAD, BREAD, etc.,—end in EAD but not all do (cf. DEAF) and not all of the words that end in EAD are exceptions to the standard EA correspondence (cf. BEAD; and the homographs READ and LEAD). Thus, the relationship between EA and its pronunciation is statistical. Similar statistical relations exist between the present and past tense forms of many of the English verbs; thus, many monosyllabic verbs with the short 'ih' vowel followed by a velar consonant (dig, swing) form the past tense by changing 'ih' to 'uh' (did-dug, swing-swung). Again, the regularity is statistical rather than deterministic (cf. sing-sang, and ring, which can be rang or ringed depending on the meaning intended).

One approach to learning mappings of this sort is to propose that they are handled by dual learning systems: one that learns the general rules and another that contains a list of the exceptions (Pinker and Prince, 1988; Coltheart et al., 1994). A different approach, first presented in the Rumelhart and McClelland (1986) model of past tense formation and the Sejnowski and Rosenberg (1987) NETtalk model for translation from spelling to sound, assumes that the entire quasi-regular system can be acquired in a single multilayer network. These systems share with the Medin and Schaffer model of category learning the property that individual items (in this case words)—especially those that occur frequently in the learner's experience—influence the response the network makes to other similar items. At the same time, they show how these effects of individual items can cumulate to produce outputs for novel items that conform to regularities that many examples share. There has been considerable debate about the adequacy of these one-process systems. The first models introduced did have some inadequacies, but recent models in both domains (MacWhinney and Leinbach, 1991; Plaut and McClelland, 1993) address the main concerns and demonstrate that a single system can be adequate to capture both the regularities and the exceptions. While it remains debatable whether the deeper aspects of language can be captured by neural network models, it seems clear, at least to this writer, that the problem of translation from streams of words to an appropriate semantic interpretation is quasi-regular (see McClelland, St. John and Taraban, 1989). Thus, it seems very likely that many of the statistical properties of neural network models will be evident in any successful model of language use and language acquisition.

To summarize, two very general and central tasks for cognitive systems—the task of interpretation and the task of learning—appear in essence to be statistical in nature. Artificial neural networks are attractive mechanisms for modeling such tasks because, as Cheng and Titterington make clear, neural networks are essentially devices that implement statistical processes. Given this, the current burgeoning of interactions between mathematical statistics and neural network research is a welcome

development for cognitive science. Such interactions will lead to a deeper understanding of the interpretation and learning tasks, and may ultimately help us to address other cognitive tasks, perhaps including creative thinking and scientific discovery, as well.

ACKNOWLEDGMENTS

Support for the preparation of this article was provided by NIMH Project Program Grant MH47566-03 and by NIMH Research Scientist Award MH00385-13.

Comment

B. D. Ripley

Bing Cheng and Mike Titterington have reviewed many of the areas of neural networks; their paper overlaps the flood of books on the subject. I also recommend Weiss and Kulikowski (1991) (Segre and Gordon, 1993, provide an informative review) and Gallant (1993) for their wider perspective and Wasserman (1993) for coverage of recent topics. My own review article, Ripley (1993a), covers this and many of the cognate areas as the authors comment. The five volumes of the NIPS proceedings (*Advances in Neural Information Processing Systems*, 1989–1993, various editors) provide a very wide-ranging overview of highly-selected papers. Much of the latest work is available electronically from the ftp archive at `archive.cis.ohio-state.edu` in directory `pub/neuroprose`.

At the time I received this paper to discuss, I had recently attended a NATO Advanced Study Institute on *From Statistics to Neural Networks* (whose proceedings will appear as Cherkassky, Friedman and Wechsler, 1994), which despite the direction of the title revealed that current thoughts in neural networks are not to subsume statistics in neural networks but vice versa. Many researchers in neural networks are becoming aware of the statistical issues in what they do and of relevant work by statisticians which encourages fruitful discussions.

Cheng and Titterington concentrate on similarities between statistical and neural network methods. I feel the differences are more revealing as they indicate room for improvement on at least one side. However, I believe the most important issues to be those of practice which are almost ignored in the paper. Before I turn to those, there are two points I wish to attempt to clarify.

B. D. Ripley is Professor of Applied Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, United Kingdom. This comment was written while on leave at the Isaac Newton Institute for Mathematical Sciences, Cambridge, United Kingdom.

1. PROJECTION-PURSUIT REGRESSION

The connection between multilayer perceptrons (MLPs) and projection-pursuit regression (PPR) is much deeper than the authors appear to suggest. Other empirical comparisons (apart from my own cited in the paper) are given by Hwang et al. (1992a,b, 1993), and Barron and Barron (1988) viewed PPR from a network viewpoint. In the authors' notation PPR is

$$y_i = w_{0i} + \sum_k \gamma_i \psi_k (x^T v_k),$$

where I have allowed for multiple outputs. An MLP with linear output units is the special case of logistic ψ_k; of course both PPRs and MLPs can be given nonlinear output units. Since we can approximate any continuous ψ_k of compact support uniformly by a step function and can approximate (nonuniformly) a step function by a logistic, we can approximate ψ_k uniformly by a sum of logistics. This fact plus the (elementary) approximation result for PPR of Diaconis and Shahshahani (1984) gives the approximation results of Cybenko and others. There is a version of Barron's L_2 result for PPR by Zhao and Atkeson (1992). (This point of view, approximating ψ_k by a simple neural net of one input, corresponds to organized weight-sharing between input-to-hidden-unit weights for groups of units, a sensible procedure in its own right.)

These results suggest that the approximation capabilities of MLPs and PPR are very similar (suggesting an affirmative partial answer to the question in Section 7). However, PPR will have an advantage when there are many inputs, only a few combinations of which are relevant, in making better use of each projection and hence fewer projections and parameters. My suspicion is that this is commonly the case.