A Comparison of Concept Identification in Human Learning and Network Learning with the Generalized Delta Rule

Michael Pazzani
The Aerospace Corporation
P.O.Box 92957
Los Angeles, CA 90009
and
UCLA AI Laboratory

Michael Dyer
UCLA AI Laboratory
3531 Boelter Hall
Los Angeles, CA 90024

Abstract

The generalized delta rule (which is also known as error back-propagation) is a significant advance over previous procedures for network learning. In this paper, we compare network learning using the generalized delta rule to human learning on two concept identification tasks:

- Relative ease of concept identification
- Generalizing from incomplete data

Introduction

The generalized delta rule for network learning has received a great deal of attention recently. The generalized delta rule is a learning procedure for associative networks which contain hidden units. It is a significant advance over previous network learning procedures which either (1) were limited to two-layer networks [23] which are incapable of solving a number of interesting problems [12], or (2) required stochastic units [9] and a large amount of computation for each learning cycle.

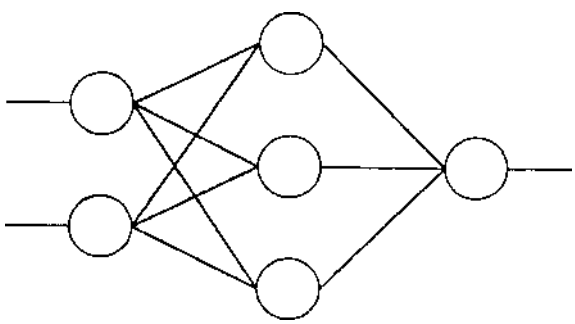

Figure 1: A multi-layer network which contains hidden units

Figure 1 presents a simple network with two input units, three hidden units, and one output unit. For each unity in a network, the output *o* given an input pattern *p* is:

$$o_{pj}=\frac{1}{1+e^{-(\sum_i w_{ji}o_{pi}+\theta_j)}}$$

where is the weight from unit *i* to *j* and θ_j is a "threshold" for unit *j*.

The generalized delta rule indicates how the weight w_{ji} on the connection from unit *i* to unit *j* should be changed after presentation of an input pattern p.

$$\Delta_p w_{ji}=\eta\delta_{pj}o_{pi}$$

where

$$\delta_{pj}=o_{pj}(1-o_{pj})(t_{pj}-o_{pj}) \quad \text{if } j \text{ is an output unit}$$

and

$$\delta_{pj}=o_{pj}(1-o_{pj})\sum_k \delta_{pk}w_{kj} \quad \text{if } j \text{ is an hidden unit}$$

where n is a parameter which controls the learning rate; t_{pj} is the target output for unit *j* with input pattern *p;* δ_{pk} is the error propagated back to unit *j* from a unit *k* whose input is o_j; and w_{kj} is the weight of the connection from unit *j* to unit *k.* The interested reader is referred to [16] for a derivation of the generalized delta rule.

The concept identification task [3] has been extensively studied in psychology. In this paper, we review a number of findings on concept identification in human subjects, and compare these findings to network learning with the generalized delta rule. In a typical concept identification experiment, a subject is shown a set of cards with different objects on them. The cards are presented to the subject one at a time in a random order and the subject is to determine whether the card belongs to the class to be learned. After each presentation, the subject is given feedback on the correctness of his response. Typically, the subject is told which attributes of the objects on the card (e.g., the number of objects, the shape of the objects, and the color of the objects) are potentially relevant. The trials continue until the subject makes no errors.

The concept identification task seems ideal for network learning. The attributes on the card are treated as input to the network. In most of the experiments, the attributes are two valued, so a binary encoding of the input is possible (e.g., for shape 1 = square and 0 = circle). The output of the network is 1 if the card is an instance of the concept and 0 otherwise. After the network classifies an input pattern, it is given feedback on the correctness of its output so that the weights on connections between units can be modified.

For all of the simulations, the network learning algorithm is simulated on a Symbolics 3600. To ensure that the algorithm has been correctly implemented, we have run it on many of the examples in [16] and obtained similar results.

The issue addressed in this paper is an evaluation of what might be called the strong PDP hypothesis: that all cognitive processes are realized directly in a homogeneous network of connected units. This hypothesis has been entertained by Churchland [5]. At the other extreme is the physical symbol system hypothesis [14]: that all cognitive processes are symbolic manipulations of the sort in logic and lisp.

> The most fundamental contribution so far of artificial intelligence and computer science to this joint enterprise has been the notion of a physical symbol system. This concept of a broad class of systems that is capable of having and manipulating symbols, yet is also realizable within our physical

> universe, has emerged from our growing experience and analysis of the computer and how to program it to perform intellectual and perceptual tasks. The notion of symbol that it defines is internal to this concept of a system. Thus, it is a hypothesis that these symbols are in fact the same symbols that we humans have and use everyday of our lives. Stated another way, the hypothesis is that humans are instances of physical symbol systems. [14]

In between these two extremes, there are a number of possible hypotheses. After presenting the results of our simulations, we shall comment on other alternative hypotheses.

Relative ease of concept identification

There have been a large number of experiments investigating the ease of learning combination of attributes. For example, Bower and Trabasso [2] have reported that concept identification by a single affirmative attribute (e.g., blue) is easier than concept identification on a conjunction of cues (e.g., blue and square). Others [3] have found that conjunctive concepts are easier than disjunctive concepts (e.g., blue or square) and that disjunctive concepts are easier than exclusive disjunctive concepts (e.g., blue or square but not blue and square) [22]. Finally, polymorphous concepts, also called m-out-of-n, (e.g., at least two of square, blue and symmetric) have been found to be more difficult than disjunctive concepts [7]. To our knowledge, there has been no comparison of polymorphous and exclusive disjunctive concepts; both are more difficult than disjunctive. Figure 2 summarizes the relative ease of acquiring concepts.

1.	Affirmation	x
2.	Conjunction	x ∧ y
3.	Disjunction	x ∨ y
4.	Exclusive Disjunction	x ⊕ y
	Polymorphous	2-out-of-3(x,y,z)

Figure 2: Relative ease of concept Identification, as determined by the number off trials required to learn the concept

We tested the generalized delta rule in a large number of different networks and conditions. In all of the tests, there were three input units, one output unit, and a number of hidden units connected to the output unit. The following parameters were varied in the trials:

- Number of hidden units: two, eight, and twenty-four.
- Connections between units: For all of the tests each input unit was connected to every hidden unit. In addition, Mi the case of the twenty-four hidden unit test, random connections between the input and hidden units were tested.
- The value of n. varied from .1 to .8 in increments of .1.

The weights W_{ijk} were set to random numbers between -5.5 and 5.5. The threshold 6, of the hidden units were also set to random values between -5.5 and 5.5 subject to the constraint that the output was never more than than .995 or less than .005 for any input combination,[1] An output value of greater than .85 was considered to be 1, and less than .15 was considered to be 0. For each condition, five concepts corresponding to one of the classes of concepts were learned from 1000 random initial conditions and the number of presentations of each input pattern was recorded.

In the conditions tested, the number of hidden units and the value of r\ did not affect the relative ordering of the ease of concept identification considerably. Results from several simulations are shown in Figure 3. For all of the simulations, If the network failed to learn after 5000 presentations of each input pattern, the simulation was terminated.

[1]The rationale for constraining the threshold value was to Increase the rate of learning. From the generalized delta rule, it it easy to see that learning is slowest when the output approaches 0 or 1.

Hidden units:	8	8	24	24
	0.2	0.6	0.2	0.6
Affirmation	449	133	97	90
Conjunction	581	211	215	174
Disjunction	560	191	229	134
Polymorphous	611	229	228	120
Exclusive Disjunction	764	324	450	291

Figure 3: Mean number of presentations of each Input pattern before correctly Identifying the concept.

There are several conclusions which can be drawn from this simulation.

- Affirmation is always easier than other classes of concepts ($p < .001$). This result is identical to the result with human subjects.
- Usually, there is no significant difference between conjunctive disjunctive, and polymorphous concepts.[2] This result differs from the findings on human subjects [22]. A bit of analysis indicates why conjunction and disjunction are equally difficult, since by interchanging Vs and 0's the disjunction becomes conjunction (i.e., in digital circuits, an or-gate in positive logic is an and-gate in negative logic). Bruner[3] has argued that disjunction is more difficult for human subjects because they find it more difficult to work with negative Instances. The generalized delta rule does not share this difficulty.
- Exclusive disjunctions were always significantly ($p < .001$) more difficult than disjunctions. This result is identical to the result with human subjects.
- The average number of presentations required for learning in all conditions is much greater than that required by human subjects. For example, in [7], with three two-valued attributes the mean number of cards presented was nine for conjunctive, twenty-eight for disjunctive, and forty for polymorphous concepts. Note that in Figure 3, the results are reported in decks of cards (i.e., presentations of all eight input patterns). The results in Figure 3 must be multiplied by eight before being compared to human performance on this task.

Redundant relevant cues

Bower and Trabasso have extensively investigated concept identification when there are redundant attributes [2]. For example, if two attributes always vary together, (i.e. squares are always blue, and blue things are always squares), then human subjects fall into three classes: those that use one of the relevant attributes (e.g.. blue), those that use the other relevant attribute (e.g.. square) and those that use both [2]. Since Bower and Trabasso were primarily concerned with determining whether or not subjects attended to both attributes, they group together those subjects who conjunctively and disjunctively combined the redundant attributes.

Encouraged by the results in the earlier simulation where the value of r\ did not alter the result, in these simulations we did not vary the value of r\ (.25). We simulated networks with 8 and 1 hidden unit(s)[3]. In this simulation there were three input attributes,

[2]Although with twenty-four hidden units and n=.6. conjunction was significantly ($p < .05$) more difficult than disjunction and polymorphy.

[3]The networks generalize better with fewer hidden units. If there a large number of hidden units, there can be one hidden unit unit which "looks for each possible input combination.

$x_t y$ and x.[4] The value of z was always the same as the value of *x*. Presentation of the four input patterns were repeated until the network would respond with 1 when *x* (and, therefore, *t)* was 1 and with 0 otherwise. After the network had learned this concept, it was presented with all eight possible input patterns. The output value of the network determined what function it had learned. For example, if the network reported 1 only when *x* was 1, then It had learned that *x* was the relevant attribute.

In this simulation, for some input patterns which were never seen the output value might not be greater than .85 (which we consider 1) or less than .15 which we consider 0. In their simulations on generalization, Rumelhart, Hinton, and Williams [16] accept a value of greater than or equal to .5 for 1, and less then .5 for 0. We also adopted this strategy. The results of these simulations are in Figure 4.

Function	8 hidden units	1 hidden unit
z	269	300
x	267	313
$x \vee z$	119	252
xz	92	279
$x\bar{y} \vee z$	91	52
$xy \vee z$	65	49
$x \vee yz$	73	55
$x \vee \bar{y}z$	85	57
$xz \vee yz$	70	42
$xz \vee \bar{y}z$	87	41
$xy \vee xz$	57	37
$x\bar{y} \vee xz$	86	29
$xy \vee \bar{y}z$	38	32
$xz \vee x\bar{y} \vee yz$	40	3
$xz \vee x\bar{y} \vee \bar{y}z$	86	3
$xz \vee xy \vee yz$	36	2

Figure 4: Distribution of concepts learned when there are redundant relevant cues.

There are sixteen possible boolean functions consistent with the four input patterns which were presented. Of these, human subjects only report the first four z, *x, xz* and *xvi* in Figure 4. in human subjects, the attribute which is not at all correlated with the output (?) is not considered relevant. The exact distribution among the four functions depends on a number factors such as the saliency of the cues. In the Trabasso and Bower experiment, 34% classified on one attribute, 51% classified on another, and 15% classified on both attributes. In network learning with eight hidden units, the rule learned to classify the concept contains the irrelevant attribute (?) slightly more than 50% of the time. With just one hidden unit, the irrelevant attribute was included more than 25% of the time.

The results of this simulation question the ability of the generalized delta rule to arrive at a reasonable generalization when some input configurations have not presented. The concept descriptions of human subjects in the redundant relevant cue experiments are simpler than those which are learned by the generalized delta rule. Occam's razor favors a simpler hypothesis over a more complex hypothesis when both are consistent with the data. Note that "simpler is defined symbolically. A distributed network which computes *x* is just a complex as one that computes $xz \vee x\bar{y} \vee \bar{y}z$.

[4]In an experiment with human subjectt, *x* might represent shape with 0 - square and 1 - circle, *y* might represent color with 0 - red and 1 = blue, and x might represent size with 0 - big and 1 - small

Discussion

In the psychology literature, the models of the concept identification task (e.g., [2,11]) are consistent with the physical symbol system hypothesis. These models postulate that subjects generate a potential concept description in an a!l-or-none fashion and then confirm (or reject) the potential concept description with future examples. When learning simple affirmative concepts, in which only one attribute is relevant (i.e., discrimination learning) the pattern of performance remains at chance for a period of time and then suddenly jumps to perfect [20]. These results are in contrast with the strong PDP hypothesis.

Some have criticized the concept identification task because the categories learned are artificial [17]. Many natural categories such as "games" do not appear to have a set of necessary and sufficient features. Instead, It is argued [24] that many concepts are polymorphous. One encouraging result of our simulation is that the polymorphous concepts are no harder for the generalized delta rule to learn than conjunctive or disjunctive concepts.[5] However, a full model of concept identification should be able to account for the acquisition of simple concepts like "square" which have necessary and sufficient features.

Others have criticized the concept identification task because the learning takes place in an artificial environment without regard to the learner's goals or prior knowledge [13,10]. For example, consider the following more realistic redundant relevant cue experiment. Someone familiar with many sports but who has never seen a game of basketball notices that there are five players with green shirts, blond hair, and various color sneakers. When one of these players has the ball, all the players run to one end of the court. Five other players have yellow shirts, black hair, and various color sneakers. When one of these players has the ball, everyone runs to the other end of the court.[6] Two opposing players collide, and are injured. Two replacements come in, one with a green shirt and black hair, the other with a yellow shirt and blond hair. The new player with the green shirt and black hair gets the ball. To which end will everyone run? An intelligent person would use his prior knowledge of sports (i.e., players on the same team wear the same color uniform) to determine that shirt color is relevant and hair color is not relevant and make the correct prediction. This Is in sharp contrast to an artificial situation in which a learner must decide whether the color or the size of a rectangle is relevant. However, the nature of the concept identification task makes no difference to the networks we have been simulating. One way to bias the saliency of attributes in network learning is to set the initial weights differently (e.g., shirt color is initially stronger than hair color). However, a simple bias would not suffice for all problems. To see this, consider the following different task: One of the players with the green shirt and blond hair also endorses hair products. He is arrested on drug charges and the company decides to find another basketball player to represent their products. In this situation, hair color may be more important than uniform color. Instead of always favoring one attribute over another, a more complex process is required which takes Into account the goals of the learner.

In our simulations, network learning with the generalized delta rule failed to exhibit a number of similarities with human learning on a number of concept identification tasks. This is in contrast with the results of network learning on other tasks, such as classical

[5]However, it should be noted that the results on the relative ease of conotpt identification assume that the learner has no prior knowledge. A theory which explains a particular combination of features facilitates teaming For example, when causal explanations are present, linearly separable categories are easier to learn than nonNnearly separable categories. The reverse is true when there is no explanation (13). It is not clear how these results could be modeled directly in network approaches to concept learning, since the generalized delta rule is not affected by the ability to construct a causal explanation.

[6]Some may recognize this as a Lakers-Celtics game.

conditioning in animals [1] and human skill learning [6]. Models such as these seem to weaken support for the strongest version of the physical symbol system hypothesis.

Conclusion

Human learning is a very complex process and It is not clear that any single rule or strategy can account for all human learning [15]. Tulving [21] has distinguished three types of human memory, each with Its own type of learning (following Rumelhart and Norman [15]) and retrieval:

- In procedural memory, which retains connections between stimuli and responses, the learning mechanism is *tuning.* Retrieval from procedural memory is by performing (i.e., acting or perceiving).
- In semantic memory, which represents knowledge of the world, the learning mechanism is called *restructuring.* Retrieval from semantic memory is called knowing.
- In episodic memory, which represents knowledge about personally experienced events, the learning mechanism is termed *accretion.* Retrieval from episodic memory Is called remembering.

The generalized delta rule seems to correspond most directly with tuning. Indeed, it has been most successful at simulating the learning of those activities of humans and animals which improve gradually over time.

We conclude that although (1) manipulating symbolic representation is not a necessary condition for intelligent behavior [8], and (2) the symbolic level is not the best level of description for some intelligent behaviors, the symbolic level is the appropriate level of description of other human behaviors. For example, in the redundant relevant cue experiment, human subjects consistently generate concept definitions which are simpler symbolically. There are two possible ways of unifying these different levels of description within the parallel distributed processing framework:

1. Look for network architectures which implement "virtual machines" which manipulate symbols [18,19]. This approach ackowledges that humans have "connectionist" hardware, but admits that (at least by adulthood) humans have built up some capabilities which are better characterized at the symbolic level.
2. Look for network architectures and learning rules which explain intelligent behaviors without reference to symbols [4]. For example, it is possible that such an architecture can follow Occam's razor without explicitly representing hypotheses as symbols and Occam's razor as a rule.

References

[1] Barto, A. and Sutton, R. Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. *Behavioral Brain Research* 4:221 -235,1982.

[2] Bower, Gordon and Trabasso, Tom. *Attention in Learning: Theory and Research.* John Wiley and Sons, New York, 1968.

[3] Bruner, J.S., Goodnow, J.J., & Austin, G.A. *A Study of Thinking.* Wiley, New York, 1956.

[4] Carpenter, G. and Grossberg, S Adaptive Resonance Theory: Stable Self-organization of Neural Recognition Codes in Responce to Arbitrary Lists of Input Patterns. In *Proceedings of the Eight Annual Conference of the Cognitive Science Society.* Amherst. Mass, 1986.

[5] Churchland, P. *Neurophilosophy: Toward a Unified Science of Mind-Brain.* MIT Press, 1986.

[6] Cohen, N., Abrams, I., Harley, W., Tabor, L. and Sejnowski, T. Skill Learning and Repetition Priming in Symmetry Detection: Parallel Studies of Human Subjects and Connectionist Models. In *Proceedings of the Eight Annual Conference of the Cognitive Science Society.* Amherst, Mass, 1986.

[7] Dennis, I., Hamptom, J. and Lea, S. New Problem in Concept Formation. *Nature* 243:463-468, May, 1973.

[8] Derthick, M. and Plaut, D. Is Distributed Connectionism Compatible with the Physical Symbol System Hypothesis? In *Proceedings of the Eight Annual Conference of the Cognitive Science Society.* Amherst, Mass, 1986.

[9] Hinton, Geoffrey, and Sejnowski Terrence. Learning and Relearning in Boltzmann Machines. In Rumelhart, David and McClelland, James (editor), *Parallel Distibuted Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations,* pages 282-317. MIT Press, 1986.

[10] Holland, J., Holyoak, K. Nisbett, R., and Thagard, P. *Induction.* MIT Press, 1986.

[11] Levine, M. Hypothesis behavior by humans during discrimination learning. *Journal of Experimental Psychology* 71:331-338, 1966.

[12] Minsky, M. and Papert, S. *Perceptrons.* MIT Press, 1969.

[13] Murphy, Gregory and Medin, Douglas. The Role of Theories in Conceptual Coherence. *Psychology Review* 92(3):289-316, 1985.

[14] Newell, A. Physical Symbol Systems. *Cognitive Science* (2), 1980.

[15] Rumelhart, D. and Norman D. Accretion, Tuning and Restructuring: Three Modes of Learning. In Cotton, J. and Klatzky, R. (editor), *Semantic Factors in Cognition,* pages 37-53. Lawrence Erlbaum Associates, Hillsdale, NJ_M 1978.

[16] Rumelhart, David, Hinton, Geoffrey, and Williams, Ronald. Learning Internal Representations by Error Propogation. In Rumelhart, David and McClelland, James (editor), *Parallel Distibuted Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations,* pages 318-362. MIT Press, 1986.

[17] Smith E.E., and Medin, D. *Categories and Concepts.* Harvard University Press, Cambridge, Ma., 1981.

[18] Touretzky, D. BoltzCons: Reconciling Connectionism with the Recursive Nature of Stacks and Trees. In *Proceedings of the Eight Annual Conference of the Cognitive Science Society.* Amherst, Mass, 1986.

[19] Touretzky, D. and Hinton, G. Symbols among the Neurons: Details of a Connectionist Inference Architecture. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence.* Morgan-Kaufmann, Los Angeles, CA, 1985.

[20] Trabasso, Tom. Stimulus emphasis and all-or-none learning on concept identification. *Journal of Experimental Psychology* 65:83-88,1963.

[21] Tulving, E. How Many Memory Systems Are There? *American Psycologist* :385-398, May, 1985.

[22] Wells, H. The effects of transfer in disjunctive concept formation. *Journal of Experimental Psychology* 65:63-69,1963.

[23] Widrow, G., and Hoff, M. Adaptive switching circuits. In *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, Part 4,* pages 96-104. 1960.

[24] Wittgenstein, L. *Philosphical Investigations.* Blakwell, 1958.