INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Suppressive Specialization:

Implications for

Generalization and Interference

In Connectionist Networks

by

Alan Lynn Franklin

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Washington

1995

Approved by 

(Chairperson of Supervisory Committee)

Program Authorized

to Offer Degree Psychology

Date 3/14/95

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to University Microfilms, 1490 Eisenhower Place, P.O. Box 975, Ann Arbor, MI 48106, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature Alan Lynn Franklin

Date 3/14/95

University of Washington

Abstract

Suppressive Specialization:
Implications for
Generalization and Interference
In Connectionist Networks

by Alan Lynn Franklin

Chairperson of the Supervisory Committee: Professor Earl Hunt
Department of Psychology

Three experiments were conducted to investigate the relationship between generalization and interference in connectionist networks. Specific modifications to the traditional error-back propagation learning algorithm are proposed to mitigate the interference that results from introducing new knowledge into previously trained networks. Data collected from 29 human subjects was compared to networks using the traditional algorithms and networks using the modified algorithm. The results indicate substantial improvements in connectionist network performance, in both tolerance to interference and improved generalization, can result from simple modifications to the learning algorithm that acknowledge previously acquired knowledge. Furthermore, these improvements result in connectionist network performances that more closely resemble the behaviors of human subjects when faced with similar tasks.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation for the assistance and guidance provided by the members of the Supervisory Committee. Special thanks need to be given to Drs. Gonzales and Miyamoto for being good council and for being sensitive to the special needs of my education. Words will never fully express my appreciation for the assistance provided to me by Dr. Hunt. Suffice it to say I am a better and wiser person because of the opportunities he has provided me.

Many thanks to each of you.

Introduction

Organisms acquire new knowledge through experience. Some knowledge is perceived directly through sensory observation (e.g. things we see, hear, smell, etc.). Other knowledge results from reasoning about these observations (e.g. the laws of physics have resulted from reasoning about physical observations). Because our experiences come to us sequentially, an important psychological issue concerns how new experiences interact with prior knowledge during the acquisition of the new knowledge. There are two possible types of interaction. Constructive interactions are beneficial and lead to extensions of our experiences that result in improved insight, understanding, and behavior. When this type of interaction occurs we say we have generalized previously learned knowledge. Destructive interactions are not beneficial and lead to reduced insight and behavior. Previously acquired knowledge becomes distorted in such a manner that the knowledge and associated behaviors become less correct or perhaps less available. Circumstances where a person was previously able to produce correct responses to input stimuli now result in less correct responses to the same stimuli. In these cases we say the acquisition of the new knowledge has retroactively interfered with previously learned knowledge. Interference can also be proactive, in that prior knowledge influences the recognition or retention of new knowledge.

Models of cognition need to acknowledge the interaction between prior knowledge and newly acquired experiences. In particular, computational models are obligated to account explicitly for the mechanism and manner of generalization and interference in information processing. This paper considers the issues of generalization and interference in one popular type of computational model of information processing. Connectionism has, for some thirty years, been proposed as a particularly important computational model of cognition. The importance of this approach to modeling cognition is derived from the intent of connectionist researchers to produce intelligent behavior from networks that are intended to mimic biological processes. Because of this explicit intent, an important issue

is "How do patterns of generalization and interference in connectionist networks compare to generalization and interference patterns in humans?". Do the similarities and/or differences in these patterns tell us anything about the legitimacy of modeling cognition through the use of connectionist networks?

This study has been designed to investigate generalization and interference in humans and to compare these behaviors to those of connectionist networks. Specific modifications to one type of connectionist network are proposed. These changes are designed to mitigate the occurrence of interference in connectionist networks. A second purpose of the study is to determine the similarity or differences in patterns of behavior (both generalization and interference) between humans and connectionist networks and to speculate on the implications of these similarities/differences to modeling cognitive behaviors.

To facilitate the discussion of these issues, this paper begins with overview discussions of generalization, interference, and connectionism. Next will follow a description of the proposed modifications to the traditional approach to connectionist networks. Following these discussions, experimental results will be presented that identify the nature of generalization and interference in humans and connectionist networks. Finally, a discussion will be presented describing the implications of these findings to cognitive psychology.

Generalization

As described earlier, generalization is the beneficial interaction between multiple learning experiences. According to the Oxford American Dictionary, generalization means to draw a general conclusion from particular instances or to bring something into general use. In psychological terms, the first part of this definition can be interpreted as noticing patterns and relationships within the input stimuli that can be used to direct a systems

response to novel, previously unseen stimuli. This interpretation suggests generalization can not exist for a single instance or experience.

In the context of this study, I consider simple behaviors to be a systems responses to input stimuli. As such, the issue of generalization involves making appropriate responses to novel or unexperienced inputs. What makes an input novel? Novel inputs must be different in some way than any previously experienced input. The difference might reflect the occurrence of a new previously unexperienced feature, or as a unique combination of previously experienced features.

For example, consider seeing an alien being for the first time. We would encode our perception of this creature according to features we would presume to be important. As interesting as this first encounter might be, our next encounter with a second alien would be even more informative. Only at that time we would be able see how the aliens were different. Then we would be able to form ideas of which features were necessary to distinguish one alien from the other. These differences would be represented as a different set of input features or as an earlier set of features with substantively different values for some or all of the features. Therefore, the key to generalization is to anticipate and recognize variations in the input stimuli and to react in some proper manner to these differences.

Although the mechanisms for generalization might be in place, single instance experiences hold no information regarding parameter variations. With single instances of experience, an organism cannot identify which of the input features might change, and therefore might require some consideration of generalization. Even if the correct features were known, single instance experiences provide no information about the range of possible values of the feature, nor do they identify the relationship between a change in stimuli and the desired change in response. To be meaningful, generalization involves developing desirable behavioral adaptations to anticipated stimuli variations. To generalize

an organism needs to:

1) identify the appropriate feature or features that will/can vary from experience to experience,

2) recognize or define what constitutes a significant variation on these parameters, and

3) map known or potential variations in the values of those features to desired variations in response behavior.

Higher order generalizations involve recognizing higher level features and patterns in the patterns of change in feature values. These higher level variations need then to be mapped to appropriate behavior response modifications. Learning and responding to feature conjunctions and other logical combinations of features would be examples of higher level generalization.

The following paragraphs discuss how various studies that have investigated generalization. The discussions are intended to show how these studies have looked at the three components of generalization. In most of cases, each study has only considered a subset of these components.

Stimulus generalization has been viewed as a tendency to react in the same way to stimuli that are similar to previously experienced stimuli (Staddon and Reid, 1990). The key points from this perspective are that generalization is in response to similar stimuli and that the reaction is in a manner that had previously proved successful. Generalization, in these terms is not the extension of one experience to another through some form of abstraction. Rather, it is the application of a previous response to a very similar stimuli. Within this notion of generalization, the objective is to transform previously unseen stimuli into a similar form that has been previously experienced such that a previously successful response could be applied to the novel stimuli. This places a premium on identifying and operationalizing the term "similar".

Staddon and Reid proposed a computational model of generalization built on work by Shepard (1958, 1987) and Ennis (1988a, 1988b). This model sought to identify a

quantitative relationship for similarity across a single psychological dimension. Shepard proposed multidimensional scaling be used to identify the minimum number of dimensions needed to judge similarity. Once these minimum dimensions were identified, Shepard proposed an exponential relationship between differences and similarity when stimuli are measured on these dimensions. Staddon and Reid tried to derive Shepard's abstract mathematical model from a model of nerve net activation. Staddon and Reid suggested Equation 1 as a means of operationalizing this relationship.

$$\delta x_i = \alpha[(x_{i-1} - x_i) + (x_{i+1} - x_i)], \quad 0 < \alpha < 0.5 \tag{1}$$

In this equation, x_i represents the current activation of a neuron. The neighboring neurons are represented by x_{i-1} and x_{i+1}. Equation 1 produces a small change that will be applied to x_i, to produce the value of x_i at the next iteration. This small change, δx_i, causes the activation value of x_i to approach a value that is between the values of the neighboring neurons. The parameter, α, influences the rate at which the activation of the neuron approaches the average value of it's neighbors. Equation 1 assumes unidimensionality, however it extends readily to the *k* dimensional case, where a unit has $2k$ nearest neighbors.

Because of the discrete nature of this equation, networks modeled using this approach are forced to use a synchronous processing approach. If Equation 1 has any genuine biological validity, it must be a finite element version of a more continuous process. Staddon and Reid state that the equation is a discrete version of Ficks first diffusion equation. Diffusion equations describe how variables change value in relationship to neighboring variables. For example, if you place a drop of colored water into a jar filled with clear water, the colored water will diffuse throughout the clear water until it is evenly distributed throughout the jar. Initially, the dye is concentrated in a

location near the point where the drop was introduced. Eventually, the dye becomes distributed throughout the jar. A diffusion equation would describe the time dependent spread of the dye through the water. A discrete version of a diffusion equation is structured such that predictions of the changes in a variable value (i.e. concentration of the dye in our water example) are made at discrete and usually uniform time increments. In the context of Equation 1, the variable that is changing is the activation of a computational unit, in response to the activations of the neighboring units. Therefore, instead of predicting the spread of dye in a jar of water, Staddon and Reid are predicting the spread of activation across a network of interconnected computational units. The net input to any one unit at time t is equal to $\delta x_i + S_i(t)$, where $S_i(t)$ is the stimulus input at time t. The addition of time dependent stimuli prevents the system from reaching a stable uniform level of activation across all units. When combined with Equation 1, the equation needed to calculate the next value of x_i becomes Equation 2:

$$x_i(t+1) = x_i(t)[1 - 2\alpha] + \alpha[x_{i-1}(t) + x_{i+1}(t)] + S_i(t) \qquad (2)$$

Equation 2 defines a means of spreading activation across neighboring stimuli nodes. To produce changes in a stable system of interconnected units, it is important the stimuli be significantly different than the current network activations. The stimuli needs to be different in the pattern it presents and it needs to be sufficiently large in magnitude to drive the direction and magnitude of change in unit activations.

For generalization to occur, it is just as important that the response the organism makes to the generalized stimuli be significant. By this I mean the organism needs the response to impact the environment in some observable manner and the organism needs to be able to detect this impact. The response might be correct or incorrect, but it needs to be of sufficient magnitude that the organism notices the correctness or incorrectness and has the opportunity to react to the consequence of the response.

The Staddon and Reid study can be viewed as addressing portions of the first and second generalization components. In looking for similarities, they find it necessary to consider which are important features of an experience and how those features might vary from experience to experience. However, their motivation in searching for these similarities is so they can collapse the new experience onto some previous experience and justify the use of a past response. This is in effect saying generalization is not the extrapolation of knowledge from one experience to another, but rather generalization is a mechanism of limiting the number of responses an organism needs to develop. Using similarity in this manner works to minimize differences and avoids looking at important differences and determining the relevancy of those differences.

Rock, I., Lasker, A., & Simon, J. (1969) see stimulus 'generalization' as a process of recognition. According to this view, organisms respond according to the similarity of stimuli to previously experienced stimuli, by producing responses to the novel stimuli that are like the responses made to similar but different stimuli. This article sees the perception of present novel stimuli as causing a recognition of some past stimuli. It is this association with the past stimuli that induces the generalization. From this perspective, generalization is a process of comparing and contrasting stimuli such that responses to previously experienced stimuli can be applied to novel stimuli. The issue of learning becomes the adjustment of a previous responses to the desired response and the making of appropriate associations of these differences to the detectable differences between the two input instances. This suggests the process of generalization is one of combining two or more previous experiences and the accompanying responses into an adjustment in the nature of a response. This adjustment becomes encoded in such a way that it will be recalled at appropriate times in the future so that the organism can better respond to similar stimuli in the future. The authors bring up the issue of an organisms awareness of differences in stimuli and how this poses a behavioral problem for organisms. With awareness, the

organism must now explicitly decide how to respond to the detected differences in the stimuli. It is possible, and perhaps even likely that subjects over categorize the stimuli such that variations of previously seen stimuli are recognized as simple variations, but because of the differences do not seems significant, the organism responds with the previous response. This problem of being aware of differences between experiences and formally deciding whether those differences justify a response modification directly relates to the second and third component of generalization.

Jacoby, Baker,& Brooks (1989) also looked at differences in experiences. In their study they considered episodic effects on picture identification. They were interested in determining whether subjects store specific experiences or generalizations of those experiences. They looked for the effects of differences in study processing on free recall of picture names and on generalization in picture identification. In this context, generalization means a reduction in detail through stimulus summarization. This notion of generalization matches well with that of Staddon and Reid, in that different experiences are used to produce a composite or average of the experiences and that average becomes the content of memory. By looking at issues of storage, they provide a different perspective on whether differences or similarities are more important. Does an organism use differences to suggest behavioral adaptations, or are they used to identify the most similar previous experiences so that those responses can be reused?

The possible mechanics of generalization have been investigated by Posner and Keele (1968). They considered whether generalization was accomplished by storing instances of all previous stimuli or by constructing an abstraction or prototype of all previously seen stimuli. In their studies they examined the impact of variations in the stimuli on the subjects patterns of generalization. Under circumstances where there was large variations in the input stimuli, subjects were better able to properly categorize novel stimuli with large variations from the prototype. Larger variations in the range of values

for input stimulus are likely to make it easier for organisms to identify those features that are going to vary from experience to experience (generalization components 1 and 2). Posner and Keele suggests that subjects store abstracted information about each stimuli, and that wide variations in the stimuli gets encoded as part of the prototype information. Information about the possible variations in the stimuli is an essential part of being able to detect meaningful variations in stimuli values.

Ross and Kennedy (1990) examined how the use of earlier examples promotes generalizations about problem types, thus influencing what is learned about a problem domain. From their perspective, initial learning involves repeated examination of examples in search of similarities to an existing problem. The relative success of applying the solutions used in the earlier examples tells subjects something about how examples can be abstracted and generalized. This view of generalization places more emphasis on recognizing and appreciating the relationships between previous examples, rather than on developing some metric or measure of similarity of examples. This begins to address the third component of generalization, how to develop novel (or at least customized) responses to novel stimuli.

The review of these studies has shown that research on the issues of generalization have covered many of the important characteristics of generalization. Each of the studies addresses one or two of the important generalization components. Developing a comprehensive theory of generalization is an important objective because generalization may be the mechanism used by organisms to organize and manage memory activities. In this view, recognizing and reacting to novel versus similar experiences (generalization) becomes the guiding principle for how and when to commit memory resources.

Interference

As described earlier, interference is a destructive interaction between experiences. Interference, however, is not the same as negative generalization. It is possible for an

organism to make an improper generalization without affecting the systems response to the original stimuli. Consider again our example of meeting aliens. In our encounter we discover the first alien (which happened to be green) liked to eat pineapple. However, the second alien (this one is red) didn't like pineapple. Now we meet a third alien (this one is also red). We might mistakenly assume the third alien would not like pineapple because the other red alien didn't (possibly a bad generalization). This generalization to all red aliens would give us no reason for changing our knowledge of the original green alien. On the other hand, if we allowed our experience with the first red alien to influence our future behavior toward the first green alien (i.e. cause us to believe the green alien didn't like pineapples), we would say the second experience had interfered with our knowledge gained from the first experience.

To avoid interference an organism needs to:

1) identify the appropriate feature or features that will/can vary from experience to experience,

2) recognize or define what constitutes a significant variation on these parameters,

3) map known or potential variations in the values of those features to desired variations in response behavior,

4) recognize when entirely new behavioral responses are justified, and

5) protect older behaviors when acquiring newer behaviors.

This list of requirements is identical to those for generalization, with two additional requirements. These are needed to protect past knowledge and to prevent efforts to generalize from becoming destructive. How do efforts to promote generalization influence interference? If generalization is based on the recognition of similarities and differences in the input stimuli, and interference is aggravated by similarities, then generalization and interference are likely to take on an adversarial relationship. We pursue one at the expense of the other. From this position, interference might be viewed as an overapplication of generalization. The key to developing complex learning behaviors might be to find the

appropriate and delicate balance between generalization and interference.

Interference has historically received considerable attention from psychologists. Some of this attention has been directed at studying interference as a phenomena unto itself, while other studies encountered interference as a companion effect to the primary interest of study (usually learning). Most of these studies analyzed interference by examining the conditions which maximized or minimized interference effects. In general, little direct work has been done on examining the mechanisms of interference.

Bower, Thompson-Schill, and Tulving provide a review the concepts of interference in associative memory (Bower, Thompson-Schill, and Tulving, 1994). According to this article, there are two classic views of associative interference. One view sees interference as the intrusion of one concept or association into the memory of another. In this form one association is learned (e.g. A-B) followed by a second association (e.g. A-C). The two associations have the same stimulus but different responses which creates an association conflict with the second association intruding on the knowledge of the first association. Intrusive interference is seen as a competition between alternative responses to a similar stimuli. Often the degree of intrusion is a function of the similarity of the competing memories. Associations that are fully orthogonal might be expected to escape this type of interference because no common elements would exist in the encoded associations.

Associative interference can also be viewed as a form of unlearning. Under this concept, as a second association is learned, the first association undergoes progressive weakening such that recall of the association is reduced. As the performance on the A-C pair increase, a corresponding decrease in A-B performance is interpreted as evidence of unlearning. The issues of unlearning and memory permanence are well addressed in a paper by Loftus and Loftus (1980). This paper examines several popular theories of memory permanence and recovery of failed memories. The central issue of this paper is

whether humans retain traces of all previous experiences, and that memory failures result from retrieval failures rather than the loss of memory content. Loftus and Loftus conclude that while perfect retention cannot be entirely dismissed, humans appear to selectively choose which memories to replace and which to coexist. To accomplish a complex system of selective replacement (interference) and coexistence, an organism needs to implement the 4th and 5th conditions described earlier for avoiding interference.

These two interpretations of interference (competition between coexisting memories versus replacement/unlearning of previous memories) differ primarily in regard to the state of the earlier learned knowledge. In the competition/intrusion interpretation, the knowledge of the first learned association is intact, but substantially distorted or hidden such that accurate recall is inhibited. For the unlearning interpretation, the knowledge of the first learned association is actually reduced or eliminated. These differences suggest different strategies and mechanisms might be needed to overcome the effects of interference. For intrusive interference, the strategy might be to map and understand the nature of the knowledge distortion such that the original knowledge can still be retrieved. For unlearning, the earlier knowledge needs to be protected and preserved. In essence, the two competing associations need to be prevented from interacting at all. Based on our earlier discussions of generalization, it may seem more advantageous to let the two associations interact, but somehow track the nature of the resulting distortions, such that prior knowledge remains retrievable.

Hirshman, Burns, and Kuo looked at the effects of encoding strategies on interference (Hirshman, Burns, and Kuo, 1993). They considered the implications of encoding relational information versus item specific information. Burns claims proactive interference occurs when learning the A-B, A-C combination and not for a D-B, A-C combination because subjects use difference encoding strategies for the two learning conditions. For the A-B, A-C condition, subjects are thought to focus on response

information, at the expense of relational information. This interpretation suggests interference is a failure to encode distinctive information, either because all available encoding strategies are "used up" or that the subject fails to focus on information that would successfully distinguish the competing concepts. The D-B, A-C condition allows subjects to focus on both the response information and information relating to the relationship between the stimuli and responses. This increased information availability provides the subjects with a richer encoding environment, which subsequently produces more interference tolerant memories. This view of interference directly addresses the first two conditions described earlier for avoiding interference. That is the organism needs to identify the important features that vary from experience to experience and to recognize what constitutes a meaningful variation on those features. In the A-B, A-C condition, the organism may be limiting itself to features that encode response information. For the D-B, A-C condition, the organism recognizes the importance of stimulus response relational information which allows it to better recognize meaningful variations and allows it more flexibility to encode the information.

Flexibility in information encoding may provide mechanisms for avoiding interference. However, the total avoidance of interference may not be desirable. The previous discussion has shown there to be close relationship between interference and generalization. In essence, any organism that seeks to make beneficial generalizations from experiences is making itself vulnerable to interference. The two concepts appear to be linked in such a way that the goal is to find an appropriate balance between generalizations and interference.

One of the most important properties of human learning is that it does not exhibit the Markov property. A (mathematical) system has the Markov property if future states of the system can be predicted from knowledge of the present state, without any knowledge of the history by which the system arrived at its present state. For instance, the system consisting

of a rocket moving through space has the Markov property, because, in the absence of intervention, the rocket's positions over time are predictable given knowledge of the rocket's current location and velocity. Furthermore, the rocket's reaction to any well defined intervening force is also predictable given only knowledge of the current position and velocity. You do not have to know how the rocket arrived at its current state to predict how it will react to an intervening force. All you need to know is what the current state is.

To see that human learning does not have the Markov property, consider the following thought experiment. Imagine two four year old children, A and M, both of whom are equally capable of calling their footwear "shoes". Suppose, though, that A has been brought up only in an English speaking environment, whereas M, who currently never uses Spanish, spent his initial two years in a Spanish speaking environment. We transport the two children to a Spanish speaking environment and attempt to teach them to call the things on their feet "zapatos". We would wager that M would relearn the shoes-zapatos connection much more quickly than A learned it for the first time.

Our though experiment is backed up by a great deal of evidence from the laboratory. Organisms exhibit a particular sort of non-Markovian inference. For example, the following two paired associate learning sequences are not equivalent.

Seq 1: A--B --> A--C --> A--B

Seq 2: A--D --> A--C --> A--B

In the first sequence, A is first associated with B, followed an association with C, and ending by repeating the association with B. The second sequence begins with an association between A and D, followed by associating A with C, and finally associating A with B. Both sequences have the association of A with C and A with B as the last two events in the sequence. A purely Markovian system would see the outcome of these two sequences as being entirely equivalent. An organism that has a purely Markovian approach to learning would be unable to distinguish between these two sequences. Studies on

relearning savings (Ebbinghaus, 1885/1964) suggest a savings should be expected to occur for the second learning of the pair A--B for the first sequence, when compared to the equivalent learning of the pair for sequence 2. If such a savings occurs, the transition from the pair A--C to the pair A--B is not Markovian and the learning of the A--B association is influenced by previous experience.

Next, we will consider a currently popular class of theories, connectionist networks, and investigate how they handle interference and generalization. This investigation will show that connectionist networks are susceptible to interference, in particular because they posses the Markov property.

Summary of Connectionism

We now consider how generalization and interference are handled by a modern, very active theoretical approach in psychology. Connectionist networks are complex networks composed of simple computational units where each unit is intended to be the computational equivalent of a biological neuron. Since its inception, connectionist research has sought to identify those aspects of cognition that are directly associable with elemental and biologically plausible computational units. Connectionist networks have been applied to the investigation of vision (Jacobs, Jordan, & Barto, 1991; Mel, 1991; Sandon, 1990; Sabbah, 1985), language capabilities (Waltz & Pollack, 1985), recognition memory (Ratcliff, 1990) and a broad variety of other pattern recognition/categorization applications (Rumelhart & Zipser, 1985). Connectionist research is interesting to psychology to the extent humans and other animals fall into the class of connectionist learning devices. This work has attempted to demonstrate that complex networks of simple computational units can be developed that are capable of performing complex cognitive tasks. Because these investigations have been both intriguing and difficult, researchers have chosen to accommodate several operational weaknesses inherent in connectionism through simplifying non-biologically plausible means.

Connectionist theories construct functions by passing activation from input nodes, representing the stimulus, to output nodes, representing a response to that stimuli. The basic assumption of connectionism is that the computation at each node is simple and that high level processes are achieved by the network through interactions between many simple units, rather than by any one unit. The idea of modeling complex cognitive behavior in this manner is appealing because we know that the brain produces complex behavioral mappings, but that the computations done by the brain's basic computing element, the neuron, are relatively simple. The connectionist approach is an attempt to simulate brain mappings by a mathematically defined networks of artificial neurons.

A number of network configurations have been proposed for connectionist networks, all based on the assumption that the knowledge and capabilities of networks lie in the interactions between units. Within these networks each unit is represented by a simple computational model first proposed by McCulloch and Pitts (McCulloch and Pitts, 1943). In this model, the activation of a unit is the transformed sum of inputs to the unit. Figure 1 provides a graphical representation of a typical computing unit.

A computing unit weights each input (O_i) by a parameter (ω_i) which represents the strength of connection between that unit and each input. These weighted inputs are then accumulated to form the total input (I_j). The total input is transformed into an output by a transformation function (Φ). The transformation function limits the range of the possible output from a unit (both upper and lower limits) and establishes a threshold for activation. Units produce essentially no output until the sum of the weighted inputs exceeds the threshold specified in the transformation function.

This approach to modeling individual neuronal behaviors has been under investigation for many years. Hebb first proposed similar computational models in the late

1940's and early 1950's (Hebb, 1949, 1954, 1955). Further work on the computations and mechanisms of learning were proposed by Milner (1957).

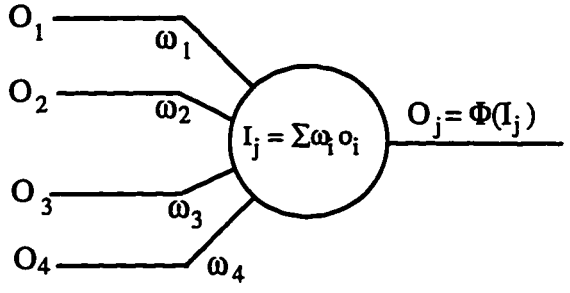

Figure 1. Typical connectionist computing unit

By combining groups of simple computational units into networks, Rosenblatt used this model to produce networks that could solve certain cognitive processing problems (Rosenblatt, 1962). In these networks, a unit receives inputs from other simple computing units in the network. The strength of each incoming signal is determined by the level of activation of the sending unit and the strength of the connection between the sending and receiving units. The receiving unit sums all of the incoming signals and transforms this total input into an appropriate activation level. This determines the level of activation of the receiving unit, which is then sent on to other units within the model according to the interconnections of the network.

There are two classic types of connectionist networks, feed forward and Hopfield networks. In feed forward networks the units are organized into layers. A layer is a group of units that operate in parallel with one another, generally receiving inputs from the same units and sending activations (outputs) to a common next layer. Activation levels propagate forward from an input layer towards the output layer. Early work by Minsky and Papert showed that simple networks of input and output units were limited to solutions that are linearly separable within the stimulus space (Minsky and Papert, 1968). As a result, there are usually one or more layers (commonly called the hidden layers) between the input and

output layers that are neither input nor output. These hidden layers enable networks to capture more complex relationships than simple input/output layer networks.

The other common class of network does not use the layering approach. In Hopfield networks the distinction between inputs and outputs is blurred. Furthermore, activation spreads throughout the network in a more asynchronous and bi-directional manner. The activations spread asynchronously because the network is not organized into layers. Instead, the spread of activation is driven by the sequence of occurrence of input stimuli. As stimuli are presented, various input units become activated. These activations then spread to whichever units are directly connected to the active units. The connections in a Hopfield network are bi-directional. Activations can be spread in either direction across any connection. Because of this bi-directionality, input units can also be used as output units and vice versa. The difference between these two types of networks results in each having its own unique characteristics of behavior. Taken together, along with any number of subtle variations, these two classes of networks form the field of connectionism.

Knowledge in a connectionist network lies in the pattern and strengths of connections between units. Therefore, the task in building connectionist networks is to determine the appropriate pattern and strengths of connections between units in the network. Rumelhart, Hinton, and Williams are credited with developing the first formal learning algorithm for feed forward networks (Rumelhart, Hinton, & Williams, 1986). While various other approaches have been developed for training (teaching) connectionist networks, the basic model of the simple computational unit remains relatively unchanged from the original concept.

In this present work, I will focus on learning and generalization in feed-forward networks. The task will be to develop feed-forward networks that acquire knowledge from sequential experiences and that generalizes appropriately from these experiences. Figure 2 shows a typical three layer feed forward network with four input units, three hidden units,

and four output units. In this study, inputs to the network are represented as binary vectors and outputs are represented as vectors of real numbers. All computational units in the network are restricted to output values between 0 and 1. The net input to the *jth* unit of a

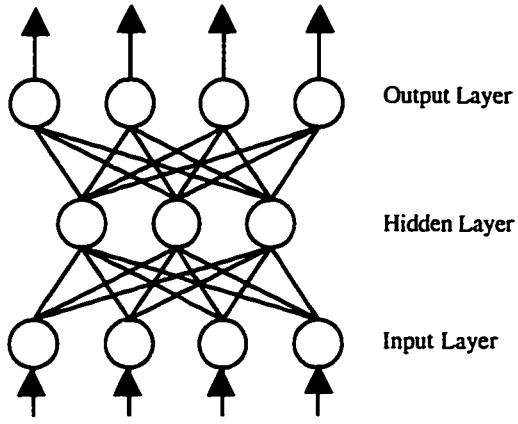

Figure 2. Typical Feed Forward Network

network is given by Equation 3. Activations from the *ith* unit of a previous layer (o_i) are weighted by the strength of connection (w_{ij}) to that unit. These weighted activations are summed for all units of the previous layer to produce a net input. This net input can range from $-\infty$ to $+\infty$, therefore a transform (or squashing) function, given by Equation 4, is used to limit the output of a unit to between 0 and 1. Learning in the network is accomplished by adjusting

$$net_j = \Sigma w_{ij} o_i \tag{3}$$

$$o_j = 1/[1 + \exp(-net_j)] \tag{4}$$

the strengths of connections for each unit within each layer such that the squared difference between the output activations and some desired output pattern (target vector) are minimized. Observation of differences between the output unit activations and the target vectors are used to generate an error signal at the output layer, that is then apportioned back across the network (error-back propagation) in a manner that reduces the magnitude of the error signal.

Given the squashing function described in Equation 4, the appropriate error signal is given by Equation 5. In this equation, t_j represents the *jth* position in the target vector and o_j is the activation or output of the *jth* output unit. The error signal, computed by Equation 5, is then apportioned to the connection strengths (weights) of the previous layer according to Equation 6. In this equation, δ_j is the error signal from Equation 5, o_i is the activation level of the *ith* unit from the previous layer, and η is a general learning parameter that influences the rate at which weights can be adjusted. Equation 6 governs

$$\delta_j = (t_j - o_j)o_j(1 - o_j) \tag{5}$$

$$\Delta w_{ij} = \eta \delta_j o_i \tag{6}$$

the modifications of the connection strengths of the output layer. To adjust the weights of the hidden layer, the error signal from the output layer, δ_j , is used in Equation 7 to compute the appropriate error signal. In the equation, δ_j is the error signal from the

$$\delta_i = o_i(1 - o_i)\Sigma w_{ij}\delta_j \tag{7}$$

previous computations and w_{ij} represents the strength of connection from the *ith* unit of the hidden layer to the *jth* unit of the output layer. This new value of δ_i is then used in the same manner as δ_j in Equation 6 to produce the desired adjustments to the connections between the hidden layer and the input layer. If more than three layers are used in the network, Equations 6 and 7 are repeated for each subsequent layer requiring adjustment.

Interference has traditionally been acknowledged in connectionism by virtue of the dependence of all networks on *sweep* learning of examples. Sweep learning refers to training connectionist networks by assembling sets of examples and presenting those examples to the network in a series of training cycles or epochs. In sweep learning, each

example in a sequence is presented one after the other, with each example contributing to some small adjustment in the knowledge of the network. The entire sequence is presented repeatedly until each member of the sequence has been learned to some desired performance criteria. For example, consider the goal of training a network to recognize four binary vectors (A, B, C, and D). To successfully train the network, the vectors would be repeatedly presented to the network as ABCDABCDABCDABC... At each presentation, the networks response to each vector would be adjusted until the network properly responds to each individual vector.

In order to avoid interference, learning within connectionist networks has been restricted to sweep learning. While sweep learning has allowed research to progress in interesting ways, the dependence on this learning approach has been subtly undermining the general faith in the entire field of connectionism. Regardless of how a network may perform, there remains the understanding that the knowledge contained in the network is fragile and can only be established through sweep training. Any attempt to incrementally add knowledge to the network, following the initial training, will undoubtedly result in the loss of most if not all previously held information. This unfortunate characteristic of connectionist networks results from not acknowledging information that may already exist in the network.

As we have seen from Equations 3-7 presented earlier, the error back propagation learning algorithm is currently structured to adjust the network connections in response to the training cases. It is presumed that once a network has modified a set of connections as the result of legitimate training, the values of the interconnection strengths (weights) are no longer arbitrary and their values in some way represent the knowledge contained in the training cases. However, current error-back propagation has no means for acknowledging connections within the network that have weights that are no longer arbitrarily determined. Without some means of detecting meaningful values of connection weights and a

companion means of adjusting strengths of connections without significantly modifying these important connections, connectionist networks will continue to learn in a typically Markovian fashion and will continue to be susceptible to catastrophic interference. Weights will be adjusted entirely in response to any current training set, without regard for any possible consideration for existing network contents (non-arbitrary connections). Until the issue of catastrophic interference has been successfully addressed, all accomplishments in connectionist research will be required to acknowledge the special conditions needed to establish and train the networks.

McLaren describes an approach for modifying the traditional error-back propagation algorithm such that the problem of catastrophic interference is avoided (McLaren, 1993). The strategy behind this approach is to allow to changes in connection strengths as a function of the status of the network. In McLaren's approach the parameter of interest is the learning rate parameter. Each hidden layer unit is allowed to adapt the learning rate applied to that unit independently of the other hidden units. McLaren's model also uses changes in the bias levels of individual units to alter the magnitude of the output for a given input signal. This has the effect of reducing the magnitude of potential connection strength modifications. (See Equation 6 and 8a).

McRae and Hetherington describe another approach for reducing catastrophic interference in connectionist networks (McRae and Hetherington, 1993). The premise of this approach is that interference occurs in unconstrained networks. If a network were limited in the number of hidden units that could represent a concept, and groups of these limited numbers of units were used to construct the hidden layer, McRae and Hetherington suggest catastrophic interference would be reduced or eliminated. In this approach, the network is constructed with specific concepts in place (pretrained). The problem now becomes how to constrain these knowledge laden units of the hidden layer as new knowledge is encountered. More importantly, what rationale is used in forming the initial

knowledge containing network. In other words, the problem has been changed from building a self-organising network to building a pre-organized network that is tolerant of subsequent exposure to experiences.

While investigating the applicability of connectionist networks to recognition memory, Ratcliff (1990) investigated a number of approaches that seemed plausible mechanisms for reducing or mitigating interference from later training of examples (adding additional cases after the initial sweep training has been completed). In his efforts, Ratcliff systematically investigated a variety of modifications to the traditional error back propagation algorithm to determine their effect on interference. His study used a multilayer network with a traditional error-back propagation approach to learning and an encoder type network (Ackley, Hinton, & Sejnowski, 1985). Encoder networks have the task of reproducing the configuration of the input vectors at the output layer. In his investigation, Ratcliff used a small encoder network (four input units, three hidden units, and four outputs, Figure 2) to demonstrate the problem of interference and a larger network to determine whether interference is a general characteristic of connectionist networks or a symptom of smaller networks. None of the models investigated at that time provided a satisfactory solution to the problem of interference and interference was observed in both the large and small networks.

Kruschke's model of category learning provides a typical example of the role interference has played in joint connectionist/psychological research (1992). Kruschke developed a example-based model of category learning. The goals of this model were to model learning by determining relevant features of a stimuli and by associating specific exemplars with categories. Kruschke's model was implemented as a feed-forward connectionist network, with three layers of computing units. The model differed from traditional feed-forward networks in that each node of the hidden layer corresponded to a particular combination of features in the total stimulus space. When a stimuli is presented

to the model, each hidden node reaches an activation that is related to the psychological similarity of the stimuli to the position in stimulus space that is represented by that node. The more similar a stimuli is to the position represented by a node, the stronger the activation for that node. Kruschke provided for a dynamic definition of similarity. As examples were presented, the network adapted to increase or decrease the importance of individual features of the stimuli. In this way, individual nodes of the hidden layer increased or decreased their relative impact on the output of the network. Kruschke viewed this change in relative importance as a change in the amount the network attended to a particular feature. Being able to model selective attention is an important step in acknowledging generalization and interference.

Kruschke's model showed promising performance characteristics regarding interference. Because the hidden layer partitioned the possible stimulus space, and each hidden layer node only responded to inputs that were similar to the space represented by that node, interference between exemplars was greatly reduced. In fact, the extent of interference in this model is dependent upon the similarity of a stimulus to previously seen stimuli and upon the resolution and amount of overlap resulting from the makeup of the hidden layer. More hidden units will give a higher resolution and the sensitivity of each unit to differences between the input and their respective positions in stimulus space influence the overlap between hidden layer nodes. These two characteristics address the first two requirements described in the introduction for an organism to avoid interference.

While Kruschke's approach eliminated the problems of catastrophic interference, it did so by predefining important differences between examples. The approach did not allow the network to detect important differences in the examples and did not allow the network to reorganize the number and relationships between hidden nodes in response to those differences.

Suppressive Specialization

The following section introduces a new concept that addresses the interference/generalization problem within the framework of feed-forward connectionist networks. It retains all of the self organizing features of traditional connectionist networks and does not require the pre-structuring constraints of Kruschke or McRae and Hetherington.

Each time a new concept is introduced into a traditional back propagation network, the learning begins without regard for previously held knowledge. Networks using the traditional algorithm do not take into account the difference between an untrained network and one that has already learned something. As a result, the knowledge stored in the network is either completely erased, or grows heavily distorted by the recent additions. For traditional back propagation, the amount of adjustments to the weights between two units in a connectionist network is a function of the error signal of the second unit and the level of activation of the first unit (Equation 5). Adjustments are applied universally to all connections within the network. These circumstances create situations of catastrophic interference (complete or near complete loss of prior knowledge). To address this problem, two new concepts, suppression and plasticity, are introduced into the general model of the computing unit. Under suppressive specialization, the amount of adjustment possible within a unit is now modulated by the amount of suppression produced by companion units within a layer and by the degree of plasticity remaining in the unit.

Suppression is produced by units within a layer competing for involvement in the representation of a concept depicted by an input vector. The more successfully a unit participates in the representation of the current input vector, the stronger a suppression signal that unit can produce. In addition, units that produce strong suppression signals are more able to resist lateral inhibition (suppression signals from other competing units). Successful participation in representing an input vector is defined as simultaneously

producing a strong output signal and a small error signal. Units that produce strong output activations and small error signals produce large suppression signals. Units that produce either large error signals or small output activations do not produce strong suppression signals. The suppression signals act as lateral inhibition to neighboring units within a common layer.

Figure 3 shows a simple network with suppression signals present in the hidden layer. In this figure, there are nine units. Units *a,b,c* are input units; *d,e,f* are hidden units; and *g,h,i* are output units. Connections between the units are shown with signs at the tips indicating the sign of the connection and the weight of the line indicating the strength of the connection. Heavy lines indicate strong connections. For example, the connection

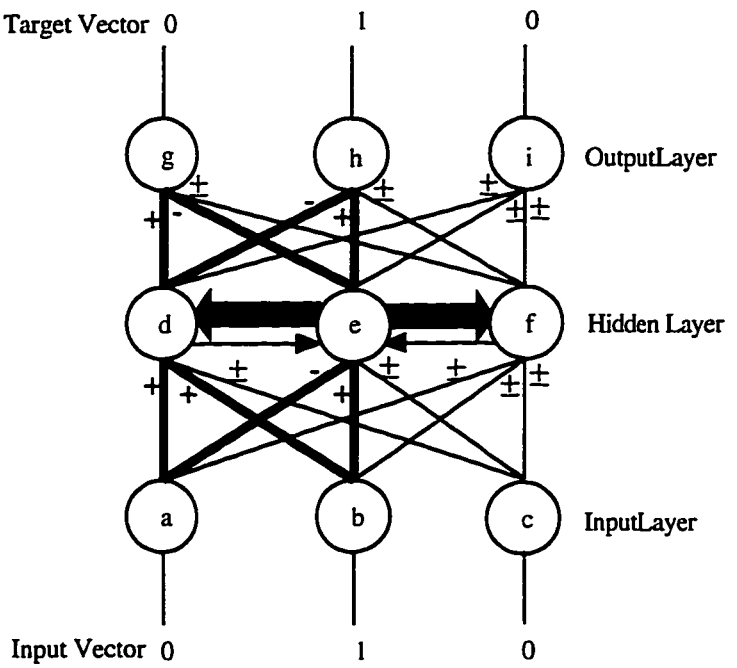

Figure 3. Example of Suppression Signals

between unit *b* and *e* is a strong positive connection. In contrast, the connection between *c* and *f* is a weak connection. The plus/minus symbol, ±, indicates the connection is some arbitrary value very near zero with either a positive or negative sign. The connections

between units *d,e,f* represent suppression signals. Unit *e* produces a large suppression signal because it produces a large output (a result of the strong connection to unit *b*) and a small error (it supports unit *h* which is properly activated). Although unit *d* is strongly activated (a strong connection to unit *b*), it produces a small suppression signal because it has a large error signal (it does not support the desired activation of unit *h*). Finally, unit *f* produces a small suppression signal because none of the connections between unit *f* and the input layer are strong enough to permit the unit to develop a strong output signal. Because unit *e* produces a strong suppression signal, it inhibits the output from the other units of the hidden layer. Therefore, even though unit *d* is producing a large output, that output is prevented from being passed on to the output layer by the suppression signal from unit *e*.

Using this approach, the units of the hidden layer self-select which units will be involved in the representation required to encode the input vector. This specialization of hidden layer involvement is accomplished through suppression of weaker units by stronger units. Because the hidden layer specializes itself, and not all units participate in the representation of concepts, the traditional fixed number of hidden units is replaced with a larger pool of hidden units. As input vectors are presented, units emerge from the pool as needed to minimize the difference between the output unit responses and the target vector. While all units in the hidden layer are generally available, not all of them will participate in the concept representation. Those units that do not activate and participate in representing an input vector remain in the pool and are available to represent subsequent input vectors. Currently suppressive specialization is applied only to the units of the hidden layer. Input units respond solely to the input vectors and output units are constrained to match the target vector. Therefore, only the units of the hidden layer are free to compete for conceptual supremacy.

As various vectors are presented to the network for training, adjustments in the plasticity are made to those units that emerge from the hidden layer pool to participate in the

representation of an input vector. A unit that has developed a strong involvement in the representation of one or more input vectors is restricted from adjusting as freely to accommodate additional vectors. This is accomplished through a reduction in the plasticity of the unit. This reduction in plasticity forces other units to emerge, as needed, from the hidden layer pool to absorb error produced by subsequently introduced input vectors. Strong representational involvement by a unit is detected by examining the strengths of connections to the previous layer. A number of strategies can be used to determine the plasticity of units. In the simplest approach, units with large values for connection weights are presumed to be highly involved in representing an earlier input vector. Therefore, the plasticity of units with strong connections is reduced which in turn reduces the amount of error that can be absorbed into these units through the adjustment of weights. Units that become highly activated (output levels very near 1.00) are also considered to be committed to previous concepts, resulting in a corresponding drop in plasticity.

Suppression and plasticity adjustment work on the hidden layer pool in opposing ways. Suppression works to restrict the number of units participating in the network representations. In other words, suppressions works to compress the pool into a few highly involved units. Plasticity reduction works to expand the number of units that emerge from the pool and become involved in concept representation. By limiting the amount of adjustment any one unit can make, reductions in plasticity force more units to emerge from the pool in order to continue reducing the error signal. The combination of these two concepts allows the network to begin with an arbitrary large number of units in the hidden layer pool while still constraining the learning (adjustment of weights) to an appropriately small number of involved units.

To incorporate suppression and plasticity adjustment into the network, a number of minor adjustments need to be made to several of the traditional error-back propagation equations. Numerical estimates of both suppression and plasticity need to be made. The

sigmoid transfer function used in Equation 4 is used to make these estimates. To generalize the sigmoid function somewhat, Equation 4 is rewritten as Equation 8.

$$o_j = 1/[1 + \exp(-\varphi(I-\theta))] \tag{8}$$

In this form, I replaces net_j. The new term, φ, influences the functions sensitivity to input. It modifies the rate of transition from 0 to 1 (slope change) as inputs range from $-\infty$ to $+\infty$. The new term, θ, introduces a bias term that allows the function to be shifted to higher or lower levels of input (bias change). The effect of each of these terms is shown in Figure 4. An increase in φ results in an increase in the slope of the transform function and a positive

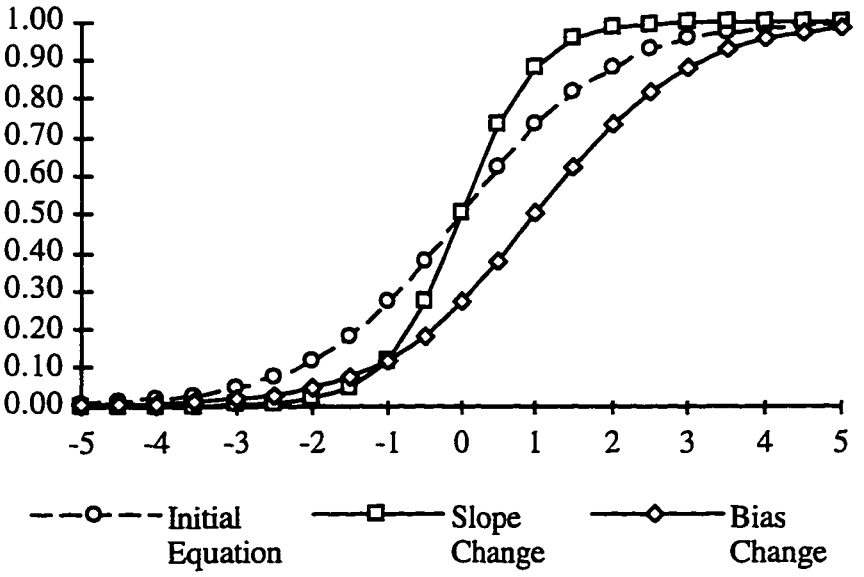

Figure 4. Modified Transform Function

value for θ causes a positive shift in the transform function.To simplify later expressions, Equation 8a will be used to represent the relationship described in Equation 8.

$$o_j = \Phi(\varphi, I, \theta) \tag{8a}$$

A network using suppressive specialization is constructed in the same manner as a

traditional network, using a hidden layer to capture non-linear relationships in the input stimuli. However, in a suppressive specialization network, the hidden layer and output layer are given slightly different operating parameters for the sigmoid transform function. For units of the output layer, φ takes on a value of 15.0 and θ takes on a constant value of 0.3. These parameter values have been empirically determined to produce desired network behaviors. In contrast, for units of the hidden layer, φ takes on a value of 22.5 and θ takes on a value that is a function of the strengths of the suppression and lateral inhibition signals described earlier. Equation 9 shows these relationships explicitly. The output of a unit, o_i , is a transform of the net input to the unit, net_i. The rate of transition from 1 to 0 and the bias required to initiate that transition are controlled by φ_4 and θ_4 , respectively. Equation 10 is used to calculate the suppression signal generated by a unit. This equation makes use of the average error, $|\varepsilon_{ave}|$, received by a unit from subsequent layers. During error back training this value is readily available by comparing output units to the target vector. However, during network evaluation, target vectors are not available. Therefore, genuine error terms can not be computed. To replace the genuine error term, an error estimate is computed. The activation level of each output unit is examined. Each unit with an activation level of 0.5 or greater is assumed to represent a target output of 1. Therefore, the error estimated for these units is $1-o_i$. Output units that produce activations of less than 0.5 are assumed to represent a target output of 0, therefore their error estimate is $-o_i$.

Equation 10 is a product of two sigmoid transfer functions. The first is sensitive to the net input, net_i, to the unit. Larger inputs produce larger transformed outputs, with the rate of transition controlled by φ_1 and θ_1. The second transform function is sensitive to the

average error assigned to a unit. For large values of error this function produces small transform values. For small error values the function produces large transform values. By multiplying the first and second transform functions together we get a composite function that has its maximum value for large net inputs and small unit errors. If either the input is small or the error is large, the composite function produces a small value.

$$o_i = \Phi(\varphi_4, net_i, \theta_4) \tag{9}$$

where $\varphi_4 = 22.5$ for hidden units,

$\varphi_4 = 15.0$ for output units,

$\theta_4 = 0.2-(V_{supp}-V_{inhibit}-1)*0.5$ for hidden units,

$\theta_4 = 0.3$ for output units.

$$V_{supp} = \Phi(\varphi_1, net_i, \theta_1)*[1-\Phi(\varphi_2, |\varepsilon_{ave}|, \theta_2)] \tag{10}$$

where $\varphi_1 = 10.0$, $\theta_1 = 0.65*(1-\rho_i)$,

$\varphi_2 = 1.0$, $\theta_2 = 0.15$, and

$|\varepsilon_{ave}|$ = the absolute value of the average error signal received by the unit.

$$\rho_i = \Phi(\varphi_3, \max(w_{ij})*o_i, \theta_3) \tag{11}$$

where $\varphi_3 = 10.0$ and $\theta_3 = 0.3$

Equation 11 provides a numerical estimate for the value of plasticity of a unit. Actually, ρ_i should be thought of in terms of rigidity. As the magnitude of the strengths of connection increase, the value of ρ_i increases. High values for ρ_i mean the ability of the unit to adjust has been reduced. The unit has become more rigid (less plastic).

The suppressive specialization strategy also modifies slightly, the manner in which

network connections are implemented. While Δw_{ij} (calculated from Equation 6) is determined in the same manner as in a traditional network, it is restricted in application. A connection (weight) can be modified by Δw_{ij} only when that modification does not weaken a committed connection or cause a connection to transition between excitatory and inhibitory in nature. A committed connection is identified as one with a current value of 0.2 or greater and belonging to a unit with $\rho > 0.5$. If either $\rho < 0.5$ or $w_{ij} < 0.2$, then the connection is allowed to be modified, as long as that modification does not change the polarity of the connection. If the adjustment of a connection strength would result in the weakening of a committed connection or cause a connection to change polarity, then that change is not applied. During training weights are updated after each presentation of an input vector/target vector pair. Weights are not adjusted during the evaluation of responses.

Experiment 1

This first experiment investigates the initial feasibility of the suppressive specialization approach to mitigating interference in connectionist networks. The investigation has been structured to replicate the findings of Ratcliff (1990) and to directly compare the performance of a network using suppressive specialization to the networks investigated in Ratcliff's work.

An important issue needs to be resolved before direct comparisons can be made to Ratcliff's results. This first experiment establishes that the implementation of the error back propagation algorithms used in this study are comparable to the algorithms used by Ratcliff. This is accomplished by training a network that is identically structured to those used by Ratcliff, but using the modified algorithm with the suppression and plasticity adjustment features disabled. Once the validity of the computational algorithms have been established, the experiment can proceed to investigate the influence of the hidden layer pool and the suppression signals and plasticity adjustments.

Methodology

Ratcliff investigated a variety of modifications to the traditional back propagation algorithm to determine their effect on interference. His study used a multilayer network with a traditional error back propagation approach to learning and the feed forward type network shown earlier in Figure 2. Ratcliff used this network to perform an encoder function. Encoder networks have the task of replicating at the output units, the exact pattern that appears on the input units. As a consequence of this function, encoder networks are constrained to have the same number of output units as they do input units.

The network used in this study is modified from Ratcliff's encoder network, to accommodate the unique self organizing characteristics of suppressive specialization. Like Ratcliff's network, this study uses four input and output units, but the three hidden units have been replaced with a large pool of 30 hidden units (Figure 5). Using suppressive specialization, it is intended that this pool self-select which hidden units specialize to the

point of representing knowledge necessary to perform the encoder task.

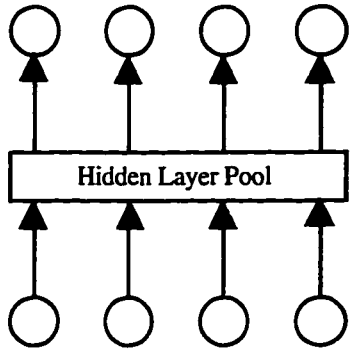

Figure 5. The multilayer encoder network with a hidden layer pool.

As with Ratcliff's study, four binary vectors of four bits each were used in this experiment. Both the traditional network and the suppressive specialization network were initially trained on three of the four input vectors, using error back propagation and a traditional sweep presentation of the training set. Following successful training of each network on the initial vectors, one additional vector was introduced to the networks. Training was performed for this vector alone, with no further presentations of the first three vectors. Following training on the fourth vector, each network was evaluated on each of the four learned vectors. In addition, two novel vectors were presented to the trained networks to determine whether either network contained any generalizations from the four training vectors.

Stimuli

The four vectors used for training in this experiment matched the four orthogonal training vectors used by Ratcliff in his 1990 study. In addition, two non-orthogonal vectors were selected for evaluation of the generalization capabilities of the networks. These six vectors (four training, two novel) are presented in Figure 6.

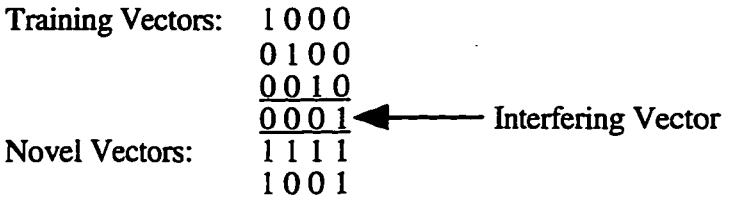

Figure 6. Training and evaluation vectors for Experiment 1.

Procedure

Each network was trained on the first three vectors until the average error between the output of the networks and the desired target values reached a value less than 0.05. Following this training, the fourth vector was presented by itself and trained into the network to the same criteria. Once training was completed, the performance of the network was evaluated across each of the four training vectors as well as the two generalization cases. In this experiment, three different networks were trained. One network was identical to Ratcliff's original encoder network, the second was a traditional connectionist network with 30 units in the hidden layer, and the third one used suppressive specialization and a pool of 30 hidden units . The second network was added to the study to determine whether differences in network performances could be the result of simply increasing the hidden layer size.

Results

The performance of these networks can be examined at two places within the experiment. The first opportunity to examine network performance occurs following training on the first three vectors. The second opportunity occurs following the training on the fourth vector. Table 1 presents the hypothetical performance of an ideal connectionist network. In this table, the column labelled "Training set" shows the four vectors used for training the netwrk. These four vectors have been grouped to show those vectors that were trained together as a set. The column labelled "Test vectors" shows the vectors that were applied after training, for the purpose of evaluating the response of the trained network.

The "Response" portion of the table shows the response of each output unit of the network. For an ideal network, the responses in the table would exactly match the test vector column. This general form of table is used to report all of the findings in this study.

Table 1

Hypothetical Performance of an

Idealized Connectionist Network

Training set	Test Vectors	Response
1000	1000	1.00 0.00 0.00 0.00
0100	0100	0.00 1.00 0.00 0.00
0010	0010	0.00 0.00 1.00 0.00
0001	1000	1.00 0.00 0.00 0.00
	0100	0.00 1.00 0.00 0.00
	0010	0.00 0.00 1.00 0.00
	0001	0.00 0.00 0.00 1.00

The performance of the traditional network, using three hidden units is presented in Table 2. These results are directly comparable to those obtained by Ratcliff, and are in essence identical to the results presented in Ratcliff's paper (1990). After training on the first three vectors, the traditional network performs at near ideal levels. However, after training on the fourth vector, a significant in the desired performance for the first three vectors and an identifiable dominance of the fourth vector can be seen. For example, for the vector {0100} the output level for the second unit (after the initial training on the first three vectors) is 0.96. After training on the fourth vector, this performance drops to 0.66 when the desired output level is 1.00. Also, the output level for the fourth unit increases from 0.01 to 0.96 for the same cases. The drop in performance for the second unit and the increase in output level for the fourth unit are evidence of retroactive interference.

Table 2

Modified Learning Algorithm with 3 Hidden Units

No Suppression or Adjustment of Plasticity

Training set	Test Vectors	Response
1000	1000	0.96 0.03 0.03 0.01
0100	0100	0.02 0.96 0.03 0.01
0010	0010	0.03 0.03 0.95 0.01
0001	1000	0.86 0.00 0.01 0.77
	0100	0.00 0.66 0.00 0.96
	0010	0.01 0.00 0.71 0.88
	0001	0.01 0.02 0.02 0.95
	1111	0.01 0.00 0.01 0.96
	0110	0.00 0.17 0.26 0.94

Table 3 presents the results of a traditional network with 30 hidden units instead of 3. This network shows the same type of performance degradation as the previous network. Again, there is a general drop in output levels for previously learned vectors and a strong tendency for the fourth unit to dominate the response to any of the input vectors. In addition to their evidence of interference, neither network appears to generalize to the two novel vectors. For example, neither network even begins to approach the ideal performance for the vector [1111]. For both networks, the response to this input is essentially dominated solely by the fourth unit.

Table 4 presents the performance of the network using the suppressive specialization learning approach. In this table we see that the reduction in output levels still occur for the first three units, following training on the fourth vector. However, the dominance of the fourth unit has been eliminated, leaving the reduced outputs of the first three units to be the dominate responses. Furthermore, the network response to the two novel vectors appears to be appropriate. While the activation levels are noticeably below ideal levels, they are easily discriminated from low activation levels that represent zeroes in

the input vector.

Table 3

Modified Learning Algorithm with 30 Hidden Units

No Suppression or Adjustment of Plasticity

Training set	Test Vectors	Response			
1000	1000	0.96	0.03	0.02	0.00
0100	0100	0.02	0.95	0.03	0.00
0010	0010	0.02	0.03	0.96	0.00
0001	1000	0.84	0.00	0.00	0.92
	0100	0.01	0.70	0.00	0.93
	0010	0.00	0.00	0.69	0.94
	0001	0.03	0.03	0.03	0.96
	1111	0.02	0.16	0.04	0.93
	0110	0.00	0.31	0.19	0.93

Table 4

Modified Learning Algorithm with 30 Hidden Units

Suppression and Adjustment of Plasticity Applied

Training set	Test Vectors	Response			
1000	1000	0.97	0.01	0.02	0.02
0100	0100	0.01	0.95	0.01	0.01
0010	0010	0.01	0.01	0.97	0.02
0001	1000	0.74	0.01	0.01	0.01
	0100	0.01	0.58	0.01	0.01
	0010	0.02	0.01	0.81	0.02
	0001	0.02	0.01	0.01	0.95
	1111	0.70	0.57	0.94	0.84
	0110	0.01	0.74	0.91	0.02

For all practical purposes, it appears that the suppressive specialization approach to

training feed forward networks has a beneficial effect on the networks resistance to interference and on the ability of the network to generalize to novel input vectors.

Experiment 2

Experiment 2 extends the findings of Experiment 1 by providing direct comparisons of connectionist networks to the performance of human subjects. To facilitate these comparisons, several adaptations to the methodology presented in Experiment 1 were necessary. In the first experiment, the four training vectors were divided into two unequal sets, with three vectors in the first set and a single vector in the second set. For this experiment, the four vectors were divided equally into two vectors per set. The motivation for this change was to provide a better training task for the human subjects. If the vectors had remained grouped as in the first experiment, then in the second part of the training task, the subjects would be asked to remember a single input/output association. It was expected that this would be a trivial task for the subjects and therefore not produce the types of circumstances that would result in interference.

The second change to the experiment was the addition of a third novel vector to be used during performance evaluation. This vector was added to complete the set of possible novel vectors and to increase the number of vectors available for final performance evaluation.

In addition to the human subject data, data was collected for both a traditional connectionist network and a network using the suppressive specialization approach. The connectionist networks operated on the binary vectors, but now divided into two groups of two vectors each. For presentation to the human subjects, the binary vectors were translated into abstract shapes. Rather than seeing stimuli as sets of four binary digits, subject viewed images consisting of four abstract shapes all enclosed within a rectangle (a banner with four abstract shapes on it). Traditionally, the inputs to a connectionist network are considered "features" of the stimuli. Their presentation order is irrelevant and not a part of the information available for training the network. To remove spatial ordering information, the abstract shapes became the features of the banner presented to the subjects. If a particular shape appears in the banner, then that feature is present. Unlike a binary

vector, an abstract feature can be present if it appears in any one of the four possible locations. Four distractor shapes were used to represent zeroes, or the absence of features. If a feature was absent, any one of the four distractor shapes could be used as a place holder in the banner. No one distractor shape was associated with any one feature. Therefore, no associations could be established between a shape and the absence of any particular feature. The distractor shapes were used so there would be no information available regarding the number of features present in any one banner presentation. When a distractor shape was needed to represent the absence of a feature, the shape was randomly chosen from the available distractor shapes. No abstract shape ever appeared more than once in any particular banner. If a distractor shape was needed, and one had already been placed in the banner, the next distractor would be chosen randomly from the remaining three distractor shapes. Figure 7 shows the abstract shapes used in this experiment. Each feature was associated with a particular category.

Subjects were asked to categorize each banner into four possible categories, according to the presence or absence of features. Any one banner could belong to more than one category, if more than one feature shape were present in the banner. For example, the each of the banners shown in Figure 8 belong to Categories 1 and 3, because both the first and third features are present each the banner. As can be seen, categorization is based on whether the features were present, and not based on the spatial position of the feature.

As stated, the four orthogonal vectors were divided into two groups of two vectors each. The first group consisted of the vectors [1000] and [0100]. The second group consisted of [0010] and [0001]. Presentation order for the two groups was counterbalanced throughout the experiment.

Procedure

Fifteen University of Washington undergraduate students participated in the experiment in exchange for course credit. Subjects were seated at individual computer

terminals in sound isolation booths. All subjects were instructed, as a group, regarding the nature of the task prior to entering the booths. The stimuli was presented on a MacIntosh Plus computer using standard black and white colors. All banners were presented one at a time, with each presentation constituting a trial. At each presentation, subjects were asked to categorize the banner into four possible categories. The four categories corresponded to

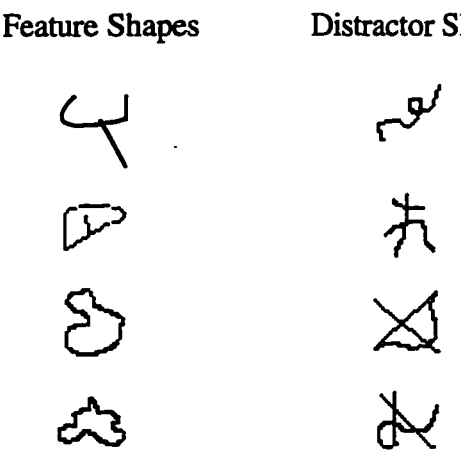

Figure 7. Abstract Feature and Distractor Shapes

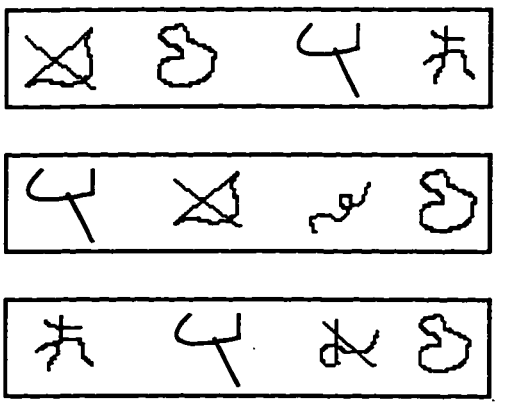

Figure 8. Three Sample Banners with Feature 1 and 3 Present

the four target features. Subjects were instructed to place each banner in as many of the four categories as possible. Each of the categories were fully independent from each other. In other words, each banner could conceivably belong to any one or any combination of the four categories.

Banners representing the target vectors were intermixed with an equal number of distractor banners. Feedback regarding the correct banner categories was presented for the target banners after the subject had made their category selections. No feedback was provided for the distractor banners. The banner and the subjects category choices remained on the screen while feedback was provided. Figure 9 shows two typical feedback displays. For the display on the left, the subject has selected Categories 2 and 4, indicated by the blackened rectangles. In this case, the selection of Category 2 is correct, but the selection of Category 4 is incorrect. This is can be identified by the small black square appearing above Rectangle 2 (a correct response) and the absence of a black square above Rectangle 4 (an incorrect response). In the display on the right, the subject has incorrectly selected Category 2 and has failed to properly select Category 1. This is indicated by the presence of a small black square above Rectangle 1 (the desired category) and the absence of a small black square above Rectangle 2 (an incorrect selection).

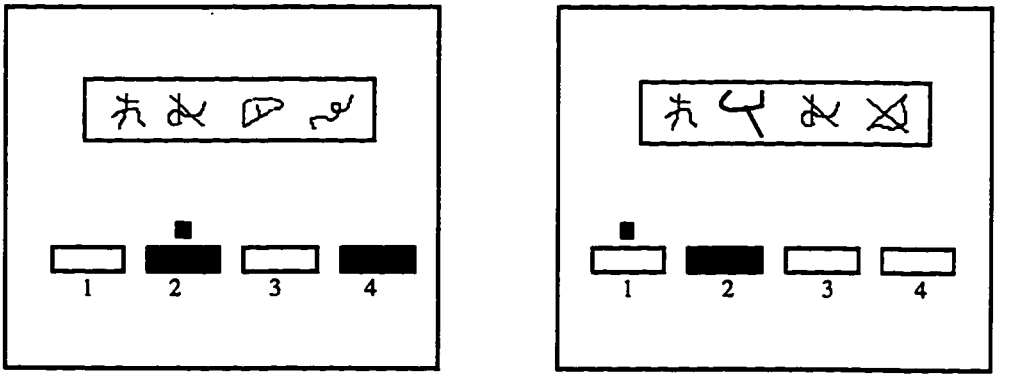

Figure 9. Typical Feedback Display

In the first part of the experiment, subjects were presented two target banners that represented vectors [1000] and [0100] and two distractor banners. For the target banners, either Feature 1 or Feature 2 were present but never both at the same time. Two distractor banners were constructed by randomly selecting four of the abstract shapes from the set of eight possible shapes. The banners were each presented one at a time. No feedback was

provided for the distractor banners. Therefore, although it would be possible for Features 1 and 2 to be present in the distractor banners, no feedback would be available to reinforce (confirm or deny) their presence. Training continued (repeated presentation of the four banners) until the subjects had successfully categorized (no inclusion or omission errors) 10 consecutive target banners. Because of the nature of the presentation of the stimuli, the probability of correctly guessing 10 consecutive banners is estimated to be approximately 0.001. Six of the fifteen subjects were dropped from the experiment because they had failed to categorize the 10 consecutive banners within 45 minutes. One additional subject was dropped because they admitted using a guessing strategy throughout the experiment. (Theoretically a one chance in 1,000,000 of successfully completing the experiment with this strategy). Data from the remaining eight subjects was used for this analysis.

Results

Results for the connectionist network closely resemble the results obtained in Experiment 1. As before, the results reported for the connectionist networks is the result of averaging over 5 separate identical networks. Table 5 shows the results for the traditional network and Table 6 shows the results for the network using suppressive specialization. The tables clearly indicate a persistence of interference across the two experiments for the traditional network and a persistent failure to generalize across the novel vectors. The system responds to the newly added third novel vector in essentially the same manner as the other two novel vectors. Furthermore, the tables show a consistent resistance to interference for the suppressive specialization network. Also, generalization has been extended to the newly added novel vector.

Table 7 provides the comparative results for the human subjects. In this data, we see some evidence of interference. Responses to both the first and second vectors indicate a lingering effect from the third and fourth features. In addition, the response to Features 1 and 2 are somewhat reduced, 0.50 and 0.38 respectively. The subjects showed a qualified

capability to generalize to the three novel stimuli. The third feature of the vector [1111] showed a response level of 0.19 (desired level was 1.00), while this same feature showed a response level of 0.19 (desired level was 0.00) for input vector [1000]. These two points represent the worst case performance for this experiment, and provide an indication of the difficulty of making statements regarding interference and generalization. If we declare the 0.19 response to be an acceptable level for responding to the vector [1111], then we are forced to accept the interference error for the vector [1000]. If, however, we declare 0.19 to be an acceptable level for a non-response to the vector [1000], then the subjects have failed to completely generalize across the vector [1111].

Table 5

Traditional Algorithm with 3 Hidden Units

Training set	Test Vectors	Response
1000	1000	0.27 0.00 0.51 0.40
0100	0100	0.00 0.21 0.48 0.56
0010	0010	0.03 0.03 0.94 0.06
0001	0001	0.03 0.03 0.07 0.93
	1111	0.01 0.01 0.42 0.57
	1001	0.16 0.00 0.12 0.83
	0110	0.00 0.14 0.88 0.15

Table 6

Suppressive Specialization Algorithm with 30 Hidden Units

Training set	Test Vectors	Response
1000	1000	0.84 0.03 0.21 0.06
0100	0100	0.03 0.81 0.23 0.05
0010	0010	0.03 0.04 0.85 0.05
0001	0001	0.04 0.02 0.05 0.88
	1111	0.85 0.75 0.94 0.92
	1001	0.84 0.02 0.21 0.89
	0110	0.05 0.86 0.89 0.07

In general, we see that the subjects performance was less than ideal, but substantially better than a traditional connectionist network. We also see a similarity in the responses of the subjects to the suppressive specialization network. For example, the third output unit has a slight tendency toward an inappropriate activation for the input vectors [1000] and [0100] (0.21 and 0.23, respectively). The human subjects also showed a slight tendency for selecting the third category when presented with the banners representing

Table 7

Human Subject Performance

Training set	Test Vectors	Response			
1000	1000	0.05	0.00	0.19	0.38
0100	0100	0.00	0.38	0.50	0.13
0010	0010	0.00	0.16	1.00	0.00
0001	0001	0.00	0.00	0.13	0.88
	1111	0.44	0.56	0.19	0.63
	1001	0.50	0.00	0.00	0.88
	0110	0.00	0.50	0.63	0.13

[1000] and [0100] (0.19 and 0.50, respectively). Furthermore,the pattern of Feature 3 and Feature 4 dominance in response to novel vectors in not seen in the human subjects. Patterns of generalized response more closely match those of the suppressive specialization network.

Experiment 3

Generalization has long been considered one of the strengths of connectionism. However, according to Staddon and Reid, generalization is a tendency to react in the same way to stimuli that are similar to previously experienced stimuli. From this position, it can be argued that in either of the first two experiments the networks (and human subjects for that matter) did not receive sufficient experience to promote generalization. Each of the training vectors used in these experiments consisted of fully orthogonal vectors with only one of the four bits active at a time. None of the experiments involved training on, or even exposure to, input stimuli that consisted of the conjunction of two or more if the input features. Without experiencing a conjunctive stimuli, it may be unreasonable to evaluate generalization across conjunctive stimuli.

Procedure

This third experiment was implemented to consider the possibility that a network must experience a conjunctive stimuli before it can create conjunctive generalizations. For this experiment, the training set of vectors was expanded to included the two conjunctive vectors, [1100] and [0011]. The new vectors were added to the training set such that the first group trained the networks on Features 1, 2, and the conjunction of 1&2, and the second group trained the networks on Features 3 & 4 in the same manner. This arrangement allows the networks to experience feature conjunctions, while still training on two features at a time.

This experiment used fourteen University of Washington undergraduates as subjects. The students participated in the experiment in exchange for psychology course credit. Banners were constructed and presented in the same manner as used in Experiment 2, with the exception of the additional banners for the vectors [1100] and [0011] that were added to the training set. Also, to maintain an approximately 50% feedback rate, two additional distractor banners were included in the training set for the experiment. These additions resulted in a total of twelve banners for the experiment. The twelve vectors were

divided into two groups with three target vectors and three distractor vectors in each group. The three target vectors were grouped so that the first group contained Features 1 and 2, and the conjunction of Features 1 and 2. The second group contained Features 3 and 4, and the conjunction of Features 3 and 4. Training on each group consisted of repeated presentations of the banners until the subjects had successfully categorized 15 consecutive target banners (5 examples of each of the three target banners). Of the initial fourteen subjects, eight failed to complete the training on the first group of banners within a time limit of 45 minutes. All of the remaining five subjects that were successful in learning the first group of banners were also successful at learning the second set of banners.

Results

Table 8, 9, and 10 show the results of this experiment for the traditional network, the suppressive specialization network, and the human subjects, respectively. The results of the

Table 8
Traditional Algorithm with 3 Hidden Units

Training set	Test Vectors	Response			
1000	1000	0.45	0.00	0.41	0.63
0100	0100	0.00	0.47	0.75	0.20
0010	0010	0.13	0.03	0.99	0.06
0001	0001	0.02	0.01	0.07	0.99
	1111	0.00	0.00	0.79	0.71
	1001	0.28	0.00	0.27	0.99
	0110	0.00	0.34	0.96	0.16

traditional network for this experiment are essentially the same as those for Experiment 2. We see a significant drop in output activation level for the first and second units following training on the second group of vectors. Additionally, we see the same pattern of third and fourth unit dominance, both in response to the training vectors and the three novel vectors.

Table 9
Suppressive Specialization Algorithm with 30 HIdden Units

Training set	Test Vectors	Response
1000	1000	0.87 0.05 0.03 0.03
0100	0100	0.03 0.81 0.02 0.02
0010	0010	0.02 0.02 0.76 0.08
0001	0001	0.03 0.02 0.04 0.89
	1111	0.73 0.77 0.68 0.78
	1001	0.69 0.02 0.06 0.66
	0110	0.03 0.72 0.57 0.19

Table 9 shows the suppressive specialization network responds in essentially the same manner as in Experiment 2. The inclusion of the conjunctive combinations of features has not appreciably affected the networks resistance to interference or the networks ability to generalize to the three novel vectors.

Table 10
Human Subject Performance

Training set	Test Vectors	Response
1000	1000	0.80 0.00 0.00 0.00
0100	0100	0.00 0.90 0.00 0.00
0010	0010	0.00 0.00 1.00 0.00
0001	0001	0.00 0.00 0.00 0.80
	1111	0.60 0.60 1.00 1.00
	1001	0.80 0.00 0.10 0.80
	0110	0.00 0.50 0.90 0.00

Unlike with the connectionist networks, the inclusion of the conjunctive vectors does seem to influence the performance of the human subjects. The principal effect is to eliminate the slight tendency to select categories 3 and 4 when presented with banners with Features 1 and 2. In addition to this slight improvement in performance, there was one

additional difference in results between Experiment 2 and Experiment 3. The number of training iterations necessary to reach the training criteria of an average error of 0.05 increased slightly when the conjunctive cases were added. Insufficient data is available to determine whether this increase is a reliable change. The change in the number of training iterations by itself is an odd curiosity. What is more intriguing is that the percentage of subjects that failed to complete the training on the first group of banners also increased with the addition of the conjunctive cases. For Experiment 2, eight out of fifteen subjects were able to complete the task, for a success rate of 53.33 percent. However, for Experiment 3, only five out of fourteen were able to complete the entire task yielding a success rate of 35.71 percent. Based on these differences and the apparent increase in the number of iterations required for the connectionist networks to reach training criterion, it appears the addition of the conjunctive cases complicates the learning of feature/category associations rather than enhances it.

Discussion

This work has provided a number of important findings regarding generalization and interference in connectionist networks. Experiment 1 provided evidence that simple changes to the basic learning algorithm of the connectionist network can result in substantial changes in the overall behavior of the network. In particular, the fundamental dependency of connectionist networks on sweep learning to prevent catastrophic interference can be avoided with relatively simple modifications to the basic learning algorithm. When using suppressive specialization these changes are designed to protect existing knowledge in the network. Furthermore, these changes (which were designed to minimize interference) also had a beneficial effect on a networks tendencies toward generalization. In addition to being to the application of connectionist networks, these findings provide a indication of a functional relationship between generalization and interference, similar to that proposed in the introduction of this paper.

Experiment 2 provided evidence that humans are subject to some of the same issues of improper generalization and interference that are commonly seen in connectionist networks. However, the patterns of generalization and interference in humans more closely matches that of a connectionist network using suppressive specialization, as opposed to the traditional error-back propagation algorithm. Responses from the first set of training vectors were reduced (failure to recognize) after subsequent training on the second set of vectors. In addition, a tendency to improperly select the second vectors when actually presented with one of the first vectors (retroactive interference) was observed. This interference was fully dominant for the network using the traditional error-back propagation algorithm. However, with the suppressive specialization algorithm and the for the human subjects, this interference was substantially smaller in magnitude.

Experiment 3 considered whether the simple cases presented in Experiment 1 & 2 had sufficient information content to support valid generalization. A possible explanation for the lack of generalization in these experiments could be that the rules of generalization,

and the allowable relationships between exemplars was not adequately represented in the training set. It is interesting to note that the addition of conjunctive information in the third experiment did have some beneficial impact on the patterns of interference/generalization for the human subjects. The incidence of retroactive interference was essentially eliminated and the generalized responses to the novel vectors showed a noticeable increase in response strength. This suggests humans go beyond reacting to and memorizing training exemplars. This evidence suggests humans specifically look for relationships within the features of the training data and react to those relationships by producing novel responses that are appropriate to the observed relationships in the stimuli. It may be that humans use this conjunctive information as a means of confirming the generality of the stimulus/response relationships they had been observing. Traditionally, connectionist networks have been interpreted as finding relationships or patterns in the training exemplars. However, this work suggests traditional connectionist networks generalize by producing approximate responses to novel stimuli, rather than observing patterns in the training exemplars and extrapolating novel responses that are adjusted from previous responses in some way that somehow matches the change in response to the observed change in input.

Rock, et al suggests generalization occurs at the time of testing, not at the time of learning (Rock et al, 1969). From a cognitive load perspective, generalizing at the time of testing will be more efficient than generalizing at the time of learning. In order to be successful at generalizing at the time of learning, an organism would need to consider all possible extrapolations of the learning example. Without some notion of the important parameters and of the range of possible parameter value variation, considering all possible generalizations would be impractical. However, if the generalizations are made at the time of testing, the organism will have important clues as to which responses might require modification. These clues come from observing differences between the features of the training exemplars and the test stimuli.

This suggestion that generalization occurs at the time of testing implies connectionist networks are not approaching the concept of learning from the proper perspective. Instead, organisms are more likely to detect differences in various stimuli and then judge how much variation in response is appropriate. This further implies that subjects have some awareness of the range of possible responses. If so, it may be that subjects judge the appropriateness of the response by considering how much of a deviation exists in the inputs (relative to the range of possible variations) and how well that variation mimics the implied variation in response (considering the perceived range of variation in the response set). This notion may also influence the search strategies used during the initial associative learning. For example, if the response set is small a strategy of trial and error followed with repeated negative confirmation trials (intentional misses) might be preferred over a systematic search and refinement.

If generalization occurs at the time of testing, and a major component of generalization is recognizing differences between previous training stimuli and current test stimuli, then several changes need to be made to the basic structure of connectionist networks. Figure 10 provides a candidate structure that may meet the requirements for connectionist generalization. This configuration has two components added to the traditional layered construction. The role of the difference discriminator is to detect meaningful differences between the current stimuli and previously experienced stimuli. Example of previously learned stimuli will be encoded in the hidden layer, perhaps in a manner similar to the hidden layer in Kruschke s ALCOVE model (Kruschke, 1992). The difference discriminator would make comparisons of the current stimuli and past exemplars, and based on the magnitudes and characteristics of those differences initiate the encoding of new exemplars in the hidden layer. A technique such as suppressive specialization might be used to incrementally add new exemplars without destroying the memory of past exemplars. The response extrapolator has the responsibility for

developing target responses and for detecting errors in those responses. Bases on input from the difference discriminator,the variations in desired response can be combinations of closely matched prior exemplars.

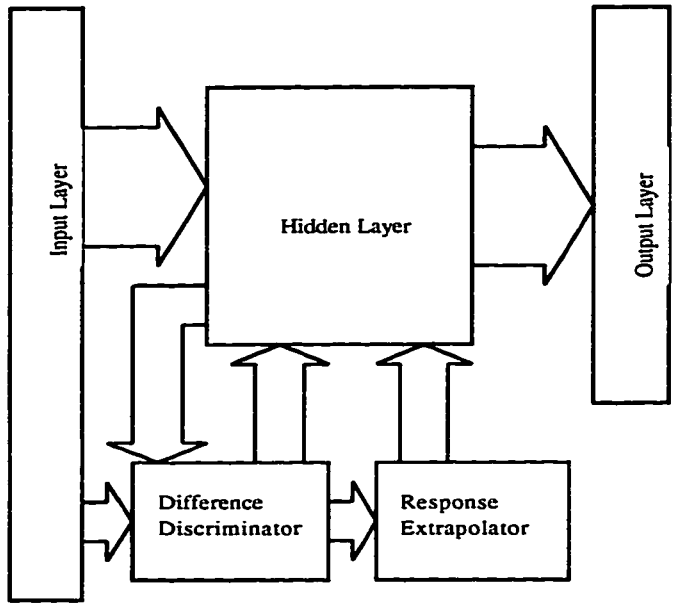

Figure 10. Candidate Connectionist Configuration

The configuration proposed in Figure 10 accounts for each of the five requirements described in the introduction that are hypothesized as being needed to avoid interference. This adjustment to the structure of a connectionist network represents a fundamental departure from the traditional approach to connectionist research in that it acknowledges the need for the network to protect existing knowledge and the need to actively search for meaningful patterns in the input stimuli, rather than simply expecting those patterns to present themselves.

References

Ackley, D.H., Hinton, G.E., & Sejnowski, T.J., 1985, A learning algorithm for Boltzman machines. Special Issue: Connectionist models and their applications, Cognitive Science, v9, 147-169.

Bower, G.H., Thompson-Schill, S., & Tulving, E., 1994, Reducing retroactive interference: An interference analysis, Journal of Experimental Psychology: Learning, Memory, and Cognition, v20, 51-66.

Ebbinghaus, H., 1885/1964 , Memory: A contribution to experimental psychology. New York: Dover.

Ennis, D.M., 1988a, Confusable and discriminable stimuli: Comment on Nosofsky (1986) and Shepard (1986), Journal of Experimental Psychology: General. v117, 408-411.

Ennis, D.M., 1988b, Toward a universal law of generalization [Comment], Science, v242, 944.

Hebb, D.O., 1949, The organization of behavior, New York: Wiley.

Hebb, D.O., 1954, The problem of consciousness and introspection, In Brain Mechanisms and Consciousness, Springfield: Thomas.

Hebb, D.O., 1955, Drives and the C.N.S. (Conceptual nervous system), Psychological Review, v62, 243-254.

Hirshman, E., Burns, D.J., & Kuo, T., 1993, Examining a processing tradeoff explanation of proactive interference, Memory & Cognition, v21, 5-10.

Jacobs, R.A., Jordan, M.I., & Barto, A.G., 1991, Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks, Cognitive Science, v15, 219-250.

Jacoby, L.L., Baker, J.G, & Brooks, L.R., 1989, Episodic effects on picture identification: Implications for theories of concept learning and theories of memory, Journal of Experimental Psychology: Learning, Memory, and Cognition, v15, 275-281.

Kruschke, J.K., 1992, ALCOVE: An exemplar-based connectionist model of category learning, Psychological Review, v99, 22-44.

Loftus, E.F and Loftus, G.R., 1980, On the permanence of stored information in the human brain, American Psychologist, v35, 409-420.

McCulloch, W.S. & Pitts, W., (1943), A logical calculus of ideas immanent in nervous activity., Bulletin of Mathematical Biophysics, v5, 115-133.

McLaren, I.P.L., 1993, APECS: A solution to the sequential learning problem, Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society, June 18-21, Institute of Cognitive Science, University of Colorado, Boulder, Colorado.

McRae, K. & Hetherington, P., 1993, Catastrophic interference is eliminated in pretrained networks, Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society, June 18-21, Institute of Cognitive Science, University of Colorado, Boulder, Colorado.

Mel, B.W., 1991, A connectionist model may shed light on neural mechanisms for visually guided reading, Journal of Cognitive Neuroscience, v3, 273-292.

Milner, P.M., 1957, The cell assembly: Mark II, Psychological Review, v64, 242-252.

Minsky, M. & Papert, S., 1968, Perceptrons, Cambridge, MA : MIT Press.

Posner, M.I. & Keele, S.W., (1968), On the genesis of abstract ideas, Journal of Experimental Psychology, v77, 353-363.

Ratcliff, R., 1990, Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions., Psychological Review, v97, 285-308.

Rock, I., Lasker, A., & Simon, J. , 1969, Stimulus generalization as a process of recognition, The American Journal of Psychology, v82, 1-22.

Rosenblatt, F., 1962, Principles of Neurodynamics; Perceptrons and the Theory of Brain Mechanisms, Spartan Books, Washington

Ross, B. H. and Kennedy, P. T., 1990, Generalizing from the use of earlier examples in

problem solving, Journal of Experimental Psychology: Learning, Memory, and Cognition, v16, 42-55.

Rumelhart, D.E., Hinton, G.E., & Williams, (1986), Learning representations by backpropagating errors, Nature, v323, 533-536.

Rumelhart, D.E. & Zipser, D., 1985, Feature discovery by competitive learning. Special Issue: Connectionist models and their applications., v9, 75-112.

Sabbah, D., 1985, Computing with connections in visual recognition of Origami objects. Special Issue: Connectionist models and their applications., v9, 25-50.

Sandon, P.A., 1990, Simulating visual attention, Journal of Cognitive Neuroscience, v2, 213-231.

Staddon, J.E.R & Reid, A.K, 1990, On the dynamics of generalization, Psychological Review, v97, 576-578.

Shepard, R.N., 1958, Stimulus and response generalization: Deduction of the generalization gradient from a trace model., Psychological Review, v65, 242-256.

Shepard, R.N., 1987, Toward a universal law of generalization for psychological science., Science, v237, 1317-1324.

Waltz, D.L. & Pollack, J.P., 1985, Massively parallel parsing: A strongly interactive model of natural language interpretation. Special Issue: Connectionist models and their applications., v9, 51-74

Curriculum Vitae

Alan Lynn Franklin

Office Address
Battelle Northwest
Box 999
Richland, WA 99352

Home Address
2513 W 40th Avenue
Kennewick, WA 99337

Personal Data — Born: July 28, 1950; Ontario, Oregon

Education

1991-1995	Program in Cognitive Psychology Department of Psychology, Seattle, WA Ph.D. (expected 3/95) Dissertation Advisor: Earl Hunt
1974-1980	MSEE, Washington StateUniversity Branch Campus, Richland, WA
1972-1974	BSEE, University of Idaho Moscow, ID
1970-1972	Engineering Program, Boise State College Boise, ID

Honors and Awards

1972	Outstanding Engineering Student Boise State College, Boise, ID

Professional Experience

1974-1989	Research Engineer Battelle Northwest, Richland, WA
1989-1991	Senior Research Scientist Battelle Northwest, Richland, WA
1994-present	Senior Research Scientist Battelle Northwest, Richland, WA

Teaching Experience

1992	Instructor, Human Performance Laboratory Department of Psychology University of Washington, Seattle, WA
1993	Teaching Assistant, Introduction to Cognitive Psychology

Department of Psychology
University of Washington, Seattle, WA

Teaching Interests
Cognition
Human Memory
Statistics and Experimental Design

General Interests
human memory
mathematical models of cognition

Presentations/Publications
available on request

Professional References

Earl Hunt, Department of Psychology
University of Washington, NI-25, Seattle, WA 98195
John M. Miyamoto, Department of Psychology
University of Washington, NI-25, Seattle, WA 98195
Richard Gonzales, Department of Psychology
University of Washington, NI-25, Seattle, WA 98195
David A. Seaver
Battelle Northwest, Richland, WA 99352