CHAPTER 5

Feature Discovery by Competitive Learning

D. E. RUMELHART and D. ZIPSER

This chapter reports the results of our studies with an unsupervised learning paradigm that we call *competitive learning*. We have examined competitive learning using both computer simulation and formal analysis and have found that when it is applied to parallel networks of neuron-like elements, many potentially useful learning tasks can be accomplished. We were attracted to competitive learning because it seems to provide a way to discover the salient, general features which can be used to classify a set of patterns. The basic components of the competitive learning scheme are:

- Start with a set of units that are all the same except for some randomly distributed parameter which makes each of them respond slightly differently to a set of input patterns.
- Limit the "strength" of each unit.
- Allow the units to compete in some way for the right to respond to a given subset of inputs.

The net result of correctly applying these three components to a learning paradigm is that individual units learn to specialize on sets of

This chapter originally appeared in *Cognitive Science*, 1985, 9, 75-112.

similar patterns and thus become "feature detectors" or "pattern classifiers." In addition to Frank Rosenblatt, whose work will be discussed below, several others have exploited competitive learning in one form or another over the years. These include von der Malsburg (1973), Grossberg (1976), Fukushima (1975), and Kohonen (1982). Our analyses differ from many of these in that we focus on the development of feature detectors rather than pattern classification. We address these issues further below.

One of the central issues in the study of the processing capacities of neuron-like elements concerns the limitations inherent in a one-level system and the difficulty of developing learning schemes for multilayered systems. Competitive learning is a scheme in which important features can be discovered at one level that a multilayer system can use to classify pattern sets which cannot be classified with a single level system.

Thirty-five years of experience have shown that getting neuron-like elements to learn some easy things is often quite straightforward, but designing systems with powerful general learning properties is a difficult problem, and the competitive learning paradigm does not change this fact. What we hope to show is that competitive learning is a powerful strategy which, when used in a variety of situations, greatly expedites some difficult tasks. Since the competitive learning paradigm has roots which go back to the very beginnings of the study of artificial learning devices, it seems reasonable to put the whole issue into historical perspective. This is even more to the point, since one of the first simple learning devices, the perceptron, caused great furor and debate, the reverberations of which are still with us.

In the beginning, thirty-five or forty years ago, it was very hard to see how anything resembling a neural network could learn at all, so any example of learning was immensely interesting. Learning was elevated to a status of great importance in those days because it was somehow uniquely associated with the properties of animal brains. After McCulloch and Pitts (1943) showed how neural-like networks could compute, the main problem then facing workers in this area was to understand how such networks could learn.

The first set of ideas that really got the enterprise going were contained in Donald Hebb's *Organization of Behavior* (1949). Before Hebb's work, it was believed that some physical change must occur in a network to support learning, but it was not clear what this change could be. Hebb proposed that a reasonable and biologically plausible change would be to strengthen the connections between elements of the network only when both the presynaptic and postsynaptic units were active simultaneously. The essence of Hebb's ideas still persists today in many learning paradigms. The details of the rules for changing weight

may be different, but the essential notion that the strength of connections between the units must change in response to some function of the correlated activity of the connected units still dominates learning models.

Hebb's ideas remained untested speculations about the nervous system until it became possible to build some form of simulated network to test learning theories. Probably the first such attempt occurred in 1951 when Dean Edmonds and Marvin Minsky built their learning machine. The flavor of this machine and the milieu in which it operated is captured in Minsky's own words which appeared in a wonderful *New Yorker* profile of him by Jeremy Bernstein (1981):

> In the summer of 1951 Dean Edmonds and I went up to Harvard and built our machine. It had three hundred tubes and a lot of motors. It needed some automatic electric clutches, which we machined ourselves. The memory of the machine was stored in the positions of its control knobs, 40 of them, and when the machine was learning, it used the clutches to adjust its own knobs. We used a surplus gyropilot from a B24 bomber to move the clutches. (p. 69)

This machine actually worked and was so fascinating to watch that Minsky remembers:

> We sort of quit science for awhile to watch the machine. We were amazed that it could have several activities going on at once in this little nervous system. Because of the random wiring it had a sort of fail safe characteristic. If one of the neurons wasn't working, it wouldn't make much difference and with nearly three hundred tubes, and the thousands of connections we had soldered there would usually be something wrong somewhere. . . . I don't think we ever debugged our machine completely, but that didn't matter. By having this crazy random design it was almost sure to work no matter how you built it. (p. 69)

In fact, the functioning of this machine apparently stimulated Minsky sufficiently to write his PhD thesis on a problem related to learning (Minsky, 1954). The whole idea must have generated rather wide interest; von Neumann, for example, was on Minsky's PhD committee and gave him encouragement. Although Minsky was perhaps the first on the scene with a learning machine, the real beginnings of meaningful neuron-like network learning can probably be traced to the work of Frank Rosenblatt, a Bronx High School of Science classmate of

Minsky's. Rosenblatt invented a class of simple neuron-like learning networks which he called perceptrons. In his book, *Principles of Neurodynamics* (1962), Rosenblatt brought together all of his results on perceptrons. In that book he gives a particularly clear description of what he thought he was doing:

> Perceptrons are not intended to serve as detailed copies of any actual nervous system. They're simplified networks, designed to permit the study of lawful relationships between the organization of a nerve net, the organization of its environment, and the "psychological" performances of which it is capable. Perceptrons might actually correspond to parts of more extended networks and biological systems; in this case, the results obtained will be directly applicable. More likely they represent extreme simplifications of the central nervous system, in which some properties are exaggerated and others suppressed. In this case, successive perturbations and refinements of the system may yield a closer approximation.
>
> The main strength of this approach is that it permits meaningful questions to be asked and answered about particular types of organizations, hypothetical memory mechanisms, and neural models. When exact analytical answers are unobtainable, experimental methods, either with digital simulation or hardware models, are employed. The model is not the terminal result, but a starting point for exploratory analysis of its behavior. (p. 28)

Rosenblatt pioneered two techniques of fundamental importance to the study of learning in neural-like networks: digital computer simulation and formal mathematical analysis, although he was not the first to simulate neural networks that could learn on digital computers (cf. Farley & Clark, 1954).

Since the paradigm of competitive learning uses concepts that appear in the work of Rosenblatt, it is worthwhile reviewing some of his ideas in this area. His most influential result was the "perceptron learning theorem" which boldly asserts:

> Given an elementary α-perceptron, a stimulus world W, and any classification $C(W)$ for which a solution exists; let all stimuli in W occur in any sequence, provided that each stimulus must reoccur in finite time; then beginning from an arbitrary initial state, an error correction procedure will always yield a solution to $C(W)$ in finite time, . . . (p. 596)

As it turned out, the real problems arose out of the phrase "for which a solution exists"—more about this later.

Less widely known is Rosenblatt's work on what he called "spontaneous learning." All network learning models require rules which tell how to present the stimuli and change the values of the weights in accordance with the response of the model. These rules can be characterized as forming a spectrum, at one end of which is learning with an error-correcting teacher, and at the other is completely spontaneous, unsupervised discovery. In between is a continuum of rules that depend on manipulating the content of the input stimulus stream to bring about learning. These intermediate rules are often referred to as "forced learning." Here we are concerned primarily with attempts to design a perceptron that would discover something interesting without a teacher because this is similar to what happens in the competitive learning case. In fact, Rosenblatt was able to build a perceptron that was able to spontaneously dichotomize a random sequence of input patterns into classes such that the members of a single class were similar to each other, and different from the members of the other class. Rosenblatt realized that any randomly initialized perceptron would have to dichotomize an arbitrary input pattern stream into a "1-set," consisting of those patterns that happened to produce a response of 1, and a "0-set," consisting of those that produced a response of 0. Of course one of these sets could be empty by chance and neither would be of much interest in general. He reasoned that if a perceptron could reinforce these sets by an appropriate rule based only on the perceptron's spontaneous response and not on a teacher's error correction, it might eventually end up with a dichotomization in which the members of each set were more like each other than like the members of the opposite set. What was the appropriate rule to use to achieve the desired dicotomization? The first rule he tried for these perceptrons, which he called C-type, was to increment weights on lines active with patterns in the 1-set, and decrement weights on lines active with patterns in the 0-set. The idea was to force a dichotomization into sets whose members were similar in the sense that they activated overlapping subsets of lines. The results were disastrous. Sooner or later all the input patterns were classified in one set. There was no dichotomy but there was stability. Once one of the sets won, it remained the victor forever.

Not to be daunted, he examined why this undesirable result occurred and realized that the problem lay in the fact that since the weights could grow without limit, the set that initially had a majority of the patterns would receive the majority of the reinforcement. This meant that weights on lines which could be activated by patterns in both sets would grow to infinite magnitudes in favor of the majority set, which in turn would lead to the capture of minority patterns by the majority set and

ultimate total victory for the majority. Even where there was initial equality between the sets, inevitable fluctuations in the random presentation of patterns would create a majority set that would then go on to win. Rosenblatt overcame this problem by introducing mechanisms to limit weight growth in such a way that the set that was to be positively reinforced at active lines would compensate the other set by giving up some weight from all its lines. He called the modified perceptrons C'. An example of a C' rule is to lower the magnitude of all weights by a fixed fraction of their current value before specifically incrementing the magnitude of some of the weights on the basis of the response to an input pattern. This type of rule had the desired result of making an equal dichotomy of patterns a stable rather than an unstable state. Patterns in each of the sets were similar to each other in the sense that they depended on similar sets of input lines to produce a response. In Rosenblatt's initial experiment, the main feature of similarity was not so much the shape of the patterns involved, but their location on the retina. That is, his system was able to spontaneously learn something about the geometry of its input line arrangement. Later, we will examine this important property of spontaneous geometry learning in considerable detail. Depending on the desired learning task, it can be either a boon or a nuisance.

Rosenblatt was extremely enthusiastic about his spontaneous learning results. In fact, his response can be described as sheer ecstasy. To see what he thought about his achievements, consider his claim (Rosenblatt, 1959):

> It seems clear that the class C' perceptron introduces a new kind of information processing automaton: For the first time, we have a machine which is capable of having original ideas. As an analogue of the biological brain, the perceptron, more precisely, the theory of statistical separability, seems to come closer to meeting the requirements of a functional explanation of the nervous system than any system previously proposed. (p. 449)

Although Rosenblatt's results were both interesting and significant, the claims implied in the above quote struck his contemporaries as unfounded. What was also significant was that Rosenblatt appeared to be saying that the type of spontaneous learning he had demonstrated was a property of perceptrons, which could not be replicated by ordinary computers. Consider the following quote from the same source:

> As a concept, it would seem that the perceptron has established, beyond doubt, the feasibility and principle of non-human systems which may embody human cognitive functions at a level far beyond that which can be achieved through present day automatons. The future of information processing devices which operate on statistical, rather then logical principles seems to be clearly indicated. (p. 449)

It is this notion of Rosenblatt's—that perceptrons are in some way superior to computers—that ignited a debate in artificial intelligence that had significant effects on the development of neural-like network models for both learning and other cognitive processes. Elements of the debate are still with us today in arguments about what the brain can do that computers can't do. There is no doubt that this was an important issue in Rosenblatt's mind, and almost certainly contributed to the acrimonious debate at that time. Consider the following statement by Rosenblatt made at the important conference on Mechanization of Thought Processes back in 1959:

> Computers seem to share two main functions with the brain: (a) Decision making, based on logical rule, and (b) control, again based on logical rules. The human brain performs these functions, together with a third: interpretation of the environment. Why do we hold interpretation of the environment to be so important? The answer, I think, is to be found in the laws of thermodynamics. A system with a completely self contained logic can never spontaneously improve its ability to organize, and to draw valid conclusions from information. (Rosenblatt, 1959, p. 423)

Clearly in some sense, Rosenblatt was saying that there were things that the brain and perceptrons, because of their statistical properties, could do which computers could not do. Now this may seem strange since Rosenblatt knew that a computer program could be written that would simulate the behavior of statistical perceptrons to any arbitrary degree of accuracy. Indeed, he was one of the pioneers in the application of digital simulation to this type of problem. What he was actually referring to is made clear when we examine the comments of other participants at the conference, such as Minsky (1959) and McCarthy (1959), who were using the symbol manipulating capabilities of the computer to directly simulate the logical processes involved in decision making, theorem proving, and other intellectual activities of this sort. Rosenblatt believed the computer used in this way would be inadequate to mimic the brain's true intellectual powers. This task, he thought, could only be accomplished if the computer or other electronic devices were used to simulate perceptrons. We can summarize these divergent

points of view by saying that Rosenblatt was concerned not only with what the brain did, but with how it did it, whereas others, such as Minsky and McCarthy, were concerned with simulating what the brain did, and didn't really care how it was done. The subsequent history of AI has shown both the successes and failures of the standard AI approach. We still have the problems today, and it's still not clear to what degree computational strategies similar to the ones used by the brain must be employed in order to simulate its performance.

In addition to producing fertilizer, as all debates do, this one also stimulated the growth of some new results on perceptrons, some of which came from Minsky. Rosenblatt had shown that a two layer perceptron could carry out any of the 2^{2^N} possible classifications of N binary inputs; that is, a solution to the classification problem had always existed in principle. This result was of no practical value however, because 2^N units were required to accomplish the task in the completely general case. Rosenblatt's approach to this problem was to use a much smaller number of units in the first layer with each unit connected to a small subset of the N inputs at random. His hope was that this would give the perceptron a high probability of learning to carry out classifications of interest. Experiments and formal analysis showed that these random devices could learn to recognize patterns to a significant degree but that they had severe limitations. Rosenblatt (1962) characterized his random perceptron as follows:

> It does not generalize well to similar forms occurring in new positions in the retinal field, and its performance in detection experiments, where a familiar figure appears against an unfamiliar background, is apt to be weak. More sophisticated psychological capabilities, which depend on the recognition of topological properties of the stimulus field, or on abstract relations between the components of a complex image, are lacking. (pp. 191-192)

Minsky and Papert worked through most of the sixties on a mathematical analysis of the computing powers of perceptrons with the goal of understanding these limitations. The results of their work are available in a book called *Perceptrons* (Minsky & Papert, 1969). The central theme of this work is that parallel recognizing elements, such as perceptrons, are beset by the same problems of scale as serial pattern recognizers. Combinatorial explosion catches you sooner or later, although sometimes in different ways in parallel than in serial. Minsky and Papert's book had a very dampening effect on the study of neuron-like networks as computational devices. Minsky has recently come to reconsider this negative effect:

> I now believe the book was overkill. . . . So after being irritated with Rosenblatt for overclaiming and diverting all those people along a false path, I started to realize that for what you get out of it — the kind of recognition it can do—it is such a simple machine that it would be astonishing if nature did not make use of it somewhere. (Bernstein, 1981, p. 103)

Perhaps the real lesson from all this is that it really is worthwhile trying to put things in perspective.

Once the problem of scale has been understood, networks of neuron-like elements are often very useful in practical problems of recognition and classification. These networks are somewhat analogous to computers, in that they won't do much unless programmed by a clever person; networks, of course, are not so much programmed as designed. The problem of finding networks of practical size to solve a particular problem is challenging because relatively small changes in network design can have very large effects on the scale of a problem. Consider networks of neuron-like units that determine the parity of their N binary inputs (see Figure 1). In the simple perceptrons studied by Minsky and Papert, units in the first layer output 1 only if all their inputs are 1 and output 0 otherwise. This takes 2^N units in the first layer, and a single linear threshold unit with a fan-in of 2^N in the second layer, to determine parity. If the units in the first layer are changed to linear threshold elements, then only N of them are required, but all must have a fan-in of N. If we allow a multilayer network to do the job, then about $3N$ units are needed, but none needs a fan-in of more than 2. The number of layers is of order $\log_2 N$. The importance of all this to the competitive learning paradigm, or any other for that matter, is that no network can learn what it is not capable of doing in principle. What any particular network can do is dependent on its structure and the computational properties of its component elements. Unfortunately, there is no canonical way to find the best network or to determine what it will learn, so the whole enterprise still has much of the flavor of an experimental science.

THE COMPETITIVE LEARNING MECHANISM

Paradigms of Learning

It is possible to classify learning mechanisms in several ways. One useful classification is in terms of the learning paradigm in which the

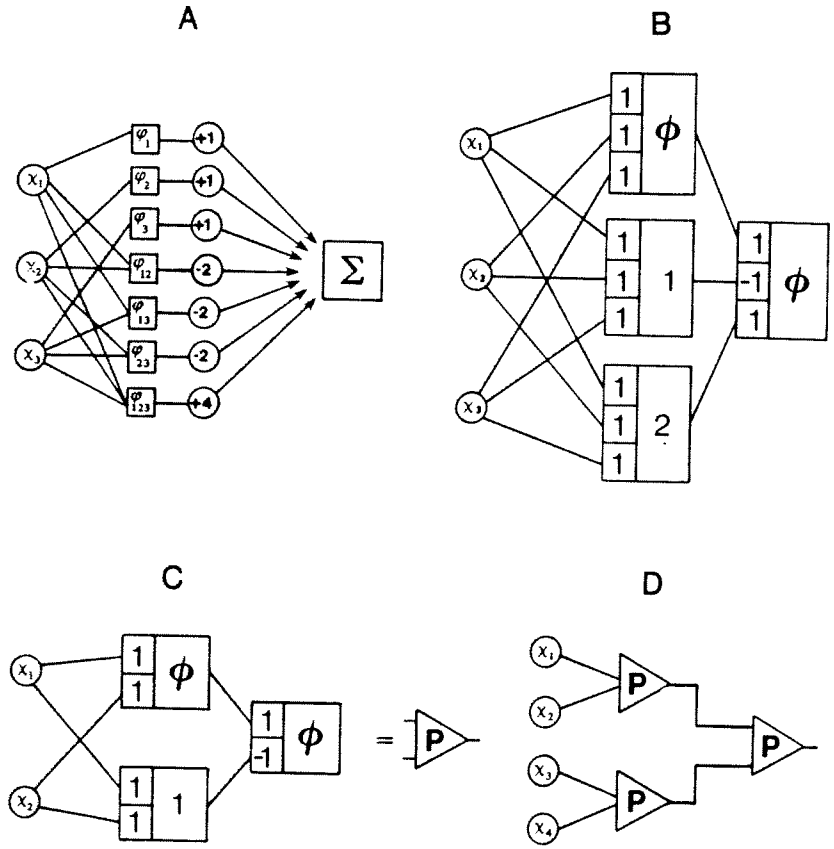

FIGURE I. *A*: Parity network from Minsky and Papert (1969). Each ϕ unit has an output of 1 only if all of its inputs are 1. Σ is a linear threshold unit with threshold of 0, i.e., like all the other linear threshold units in the figure, it fires only when the sum of its weighted inputs is greater than the threshold. This and all the other networks signal odd parity with a 1 in the rightmost unit of the network. *B*: Parity network made from two layers of linear threshold units. *C*: Three-unit network for determining the parity of a pair of inputs. *D*: Two-layer network using the subnetwork described in *(C)*. In general, the number of *P*-units is of order N and the number of layers is of order $\log_2 N$.

model is supposed to work. There are at least four common learning paradigms in neural-like processing systems:

- *Auto Associator.* In this paradigm a set of patterns are repeatedly presented and the system is supposed to "store" the patterns. Then, later, parts of one of the original patterns or possibly a pattern similar to one of the original patterns is presented, and the task is to "retrieve" the original pattern through a kind of pattern completion procedure. This is an auto-association process in which a pattern is associated with itself so that a degraded version of the original pattern can act as a retrieval cue.

- *Pattern Associator.* This paradigm is really a variant on the auto-association paradigm. A set of *pairs* of patterns are repeatedly presented. The system is to learn that when one member of the pair is presented it is supposed to produce the other. In this paradigm one seeks a mechanism in which an essentially arbitrary set of input patterns can be paired with an arbitrary set of output patterns.

- *Classification Paradigm.* The classification paradigm also can be considered as a variant on the previous learning paradigms, although the goals are sufficiently different and it is sufficiently common that it deserves separate mention. In this case, there is a fixed set of categories into which the stimulus patterns are to be classified. There is a training session in which the system is presented with the stimulus patterns along with the categories to which each stimulus belongs. The goal is to learn to correctly classify the stimuli so that in the future when a particular stimulus or a slightly distorted version of one of the stimuli is presented, the system will classify it properly. This is the typical paradigm in which the perceptron is designed to operate and in which the perceptron convergence theorem is proved.

- *Regularity Detector.* In this paradigm there is a population of stimulus patterns and each stimulus pattern, S_k, is presented with some probability p_k. The system is supposed to *discover* statistically salient features of the input *population*. Unlike the classification paradigm, there is no a priori set of categories into which the patterns are to be classified; rather, the system must develop its own featural representation of the input stimuli which captures the most salient features of the population of input patterns.

Competitive learning is a mechanism well-suited for regularity detection, as in the environment described in above.

Competitive Learning

The architecture of a competitive learning system (illustrated in Figure 2) is a common one. It consists of a set of hierarchically layered units in which each layer connects, via excitatory connections, with the layer immediately above it. In the most general case, each unit of a layer receives an input from each unit of the layer immediately below and projects output to each unit in the layer immediately above it. Moreover, within a layer, the units are broken into a set of inhibitory clusters in which all elements within a cluster inhibit all other elements in the cluster. Thus the elements within a cluster at one level compete with one another to respond to the pattern appearing on the layer below. The more strongly any particular unit responds to an incoming stimulus, the more it shuts down the other members of its cluster.

There are many variations on the competitive learning theme. A number of researchers have developed variants of competitive learning mechanisms and a number of results already exist in the literature. We have already mentioned the pioneering work of Rosenblatt. In addition, von der Malsburg (1973), Fukushima (1975), and Grossberg (1976), among others, have developed models which are competitive learning models, or which have many properties in common with competitive learning. We believe that the essential properties of the competitive learning mechanism are quite general. However, for the sake of concreteness, in this paper we have chosen to study, in some detail, the simplest of the systems which seem to be representative of the essential characteristics of competitive learning. Thus, the system we have analyzed has much in common with the previous work, but wherever possible we have simplified our assumptions. The system that we have studied most is described below:

- The units in a given layer are broken into a set of nonoverlapping clusters. Each unit within a cluster inhibits every other unit within a cluster. The clusters are winner-take-all, such that the unit receiving the largest input achieves its maximum value while all other units in the cluster are pushed to their minimum value.[1] We have arbitrarily set the maximum value to 1 and the minimum value to 0.

[1] A simple circuit for achieving this result is attained by having each unit activate itself and inhibit its neighbors. Grossberg (1976) employs just such a network to *choose* the maximum value of a set of units.

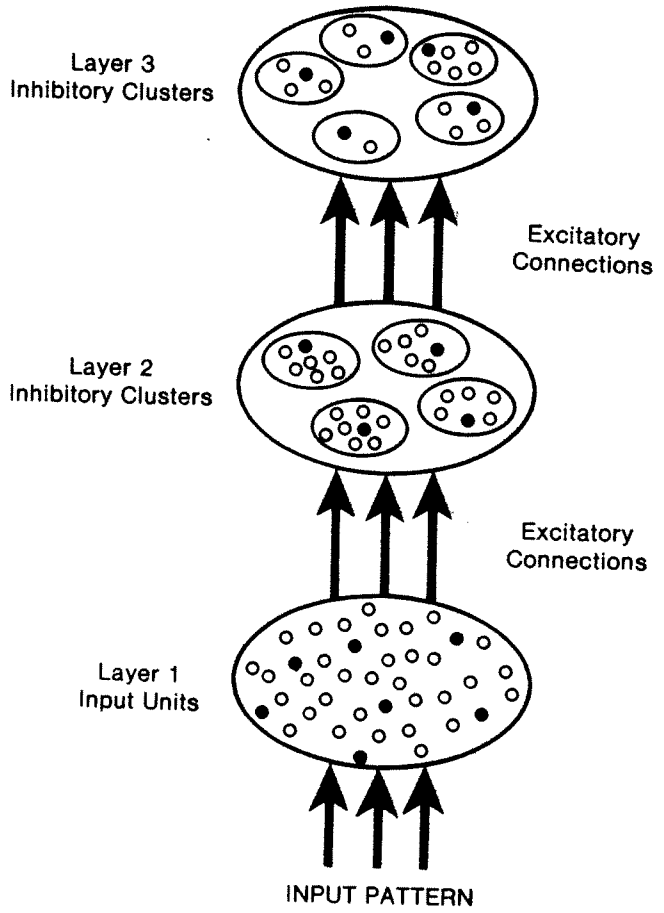

FIGURE 2. The architecture of the competitive learning mechanism. Competitive learning takes place in a context of sets of hierarchically layered units. Units are represented in the diagram as dots. Units may be active or inactive. Active units are represented by filled dots, inactive ones by open dots. In general, a unit in a given layer can receive inputs from all of the units in the next lower layer and can project outputs to all of the units in the next higher layer. Connections between layers are excitatory and connections within layers are inhibitory. Each layer consists of a set of clusters of mutually inhibitory units. The units within a cluster inhibit one another in such a way that only one unit per cluster may be active. We think of the configuration of active units on any given layer as representing the input pattern for the next higher level. There can be an arbitrary number of such layers. A given cluster contains a fixed number of units, but different clusters can have different numbers of units.

- Every element in every cluster receives inputs from the same lines.

- A unit learns if and only if it wins the competition with other units in its cluster.

- A stimulus pattern S_j consists of a binary pattern in which each element of the pattern is either *active* or *inactive*. An active element is assigned the value 1 and an inactive element is assigned the value 0.

- Each unit has a fixed amount of weight (all weights are positive) which is distributed among its input lines. The weight on the line connecting unit i on the lower (or input) layer to unit j on the upper layer, is designated w_{ij}. The fixed total amount of weight for unit j is designated $\sum_i w_{ij} = 1$. A unit learns by shifting weight from its inactive to its active input lines. If a unit does not respond to a particular pattern, no learning takes place in that unit. If a unit wins the competition, then each of its input lines give up some proportion g of its weight and that weight is then distributed equally among the active input lines.[2] More formally, the learning rule we have studied is:

$$\Delta w_{ij} = \begin{cases} 0 & \text{if unit } j \text{ loses on stimulus } k \\ g\dfrac{c_{ik}}{n_k} - gw_{ij} & \text{if unit } j \text{ wins on stimulus } k \end{cases}$$

where c_{ik} is equal to 1 if in stimulus pattern S_k, unit i in the lower layer is active and zero otherwise, and n_k is the number of active units in pattern S_k (thus $n_k = \sum_i c_{ik}$).

Figure 3 illustrates a useful geometric analogy to this system. We can consider each stimulus pattern as a vector. If all patterns contain the same number of active lines, then all vectors are the same length and each can be viewed as a point on an N-dimensional hypersphere,

[2] This learning rule was proposed by von der Malsburg (1973). As Grossberg (1976) points out, renormalization of the weights is not necessary. The same result can be obtained by normalizing the input patterns and then assuming that the weights approach the values on the input lines. Normalizing weights is simpler to implement than normalizing patterns, so we chose that option. For most of our experiments, however, it does not matter which of these two rules we chose since all patterns were of the same magnitude.

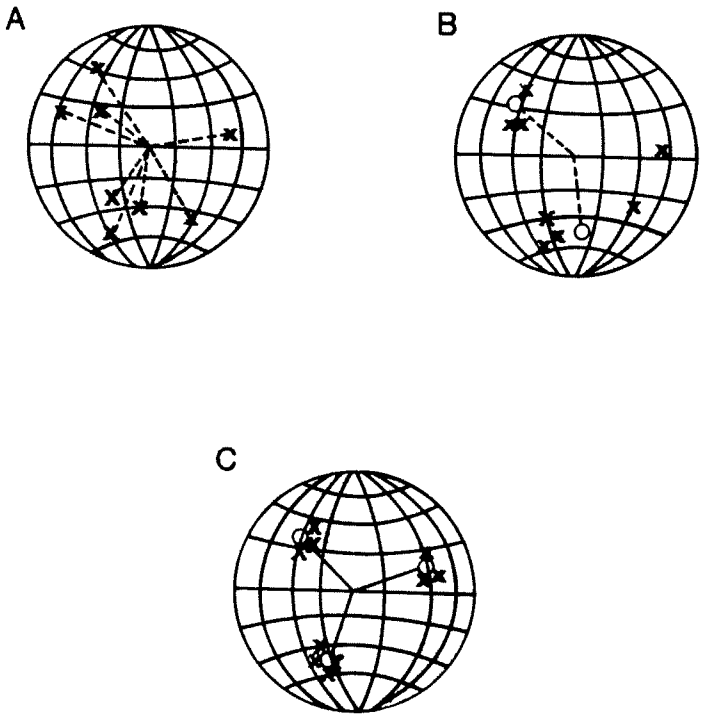

FIGURE 3. A geometric interpretation of competitive learning. *A*: It is useful to conceptualize stimulus patterns as vectors whose tips all lie on the surface of a hypersphere. We can then directly see the similarity among stimulus patterns as distance between the points on the sphere. In the figure, a stimulus pattern is represented as an ×. The figure represents a population of eight stimulus patterns. There are two clusters of three patterns and two stimulus patterns which are rather distinct from the others. *B*: It is also useful to represent the weights of units as vectors falling on the surface of the same hypersphere. Weight vectors are represented in the figure as ○'s. The figure illustrates the weights of two units falling on rather different parts of the sphere. The response rule of this model is equivalent to the rule that whenever a stimulus pattern is presented, the unit whose weight vector is closest to that stimulus pattern on the sphere wins the competition. In the figure, one unit would respond to the cluster in the northern hemisphere and the other unit would respond to the rest of the stimulus patterns. *C*: The learning rule of this model is roughly equivalent to the rule that whenever a unit wins the competition (i.e., is closest to the stimulus pattern), that weight vector is moved toward the presented stimulus. The figure shows a case in which there are three units in the cluster and three natural groupings of the stimulus patterns. In this case, the weight vectors for the three units will each migrate toward one of the stimulus groups.

where N is the number of units in the lower level, and therefore, also the number of input lines received by each unit in the upper level.

Each × in Figure 3A represents a particular pattern. Those patterns that are very similar are near one another on the sphere; those that are very different will be far from one another on the sphere. Now note that since there are N input lines to each unit in the upper layer, its weights can also be considered a vector in N-dimensional space. Since all units have the same total quantity of weight, we have N-dimensional vectors of approximately fixed length for each unit in the cluster.[3] Thus, properly scaled, the weights themselves form a set of vectors which (approximately) fall on the surface of the same hypersphere. In Figure 3B, the ○'s represent the weights of two units superimposed on the same sphere with the stimulus patterns. Now, whenever a stimulus pattern is presented, the unit which responds most strongly is simply the one whose weight vector is nearest that for the stimulus. The learning rule specifies that whenever a unit wins a competition for a stimulus pattern, it moves a percentage g of the way from its current location toward the location of the stimulus pattern on the hypersphere. Now, suppose that the input patterns fell into some number, M, "natural" groupings. Further, suppose that an inhibitory cluster receiving inputs from these stimuli contained exactly M units (as in Figure 3C). After sufficient training, and assuming that the stimulus groupings are sufficiently distinct, we expect to find one of the vectors for the M units placed roughly in the center of each of the stimulus groupings. In this case, the units have come to detect the grouping to which the input patterns belong. In this sense, they have "discovered" the structure of the input pattern sets.

Some Features of Competitive Learning

There are several characteristics of a competitive learning mechanism that make it an interesting candidate for further study, for example:

- Each cluster classifies the stimulus set into M groups, one for each unit in the cluster. Each of the units captures roughly an equal number of stimulus patterns. It is possible to consider a cluster as forming an M-ary feature in which every stimulus pattern is classified as having exactly one of the M possible values of this feature. Thus, a cluster containing 2 units acts as a binary feature detector. One element of the cluster responds when a particular feature is present in the stimulus pattern, otherwise the other element responds.

- If there is *structure* in the stimulus patterns, the units will break up the patterns along structurally relevant lines. Roughly speaking, this means that the system will find clusters if they are there. (A key problem, which we address below, is specifying the *nature* of the structure that this system discovers.)

- If the stimuli are highly structured, the classifications are highly stable. If the stimuli are less well-structured, the classifications are more variable, and a given stimulus pattern will be responded to first by one and then by another member of the cluster. In our experiments, we started the weight vectors in random directions and presented the stimuli randomly. In this case, there is rapid movement as the system reaches a relatively stable configuration (such as one with a unit roughly in the center of each cluster of stimulus patterns). These configurations can be more or less stable. For example, if the stimulus points don't actually fall into nice clusters, then the configurations will be relatively unstable, and the presentation of each stimulus will modify the pattern of responding so that the system will undergo continual evolution. On the other hand, if the stimulus patterns fall rather nicely into clusters, then the system will become very stable in the sense that the same units will always respond to the same stimuli.[4]

- The particular grouping done by a particular cluster depends on the starting value of the weights and the sequence of stimulus patterns actually presented. A large number of clusters, each receiving inputs from the same input lines can, in general, classify the inputs into a large number of different groupings, or alternatively, discover a variety of independent features present in the stimulus population. This can provide a kind of coarse coding of the stimulus patterns.[5]

[3] It should be noted that this geometric interpretation is only approximate. We have used the constraint that $\sum_i w_{ij} = 1$ rather than the constraint that $\sum_i w_{ij}^2 = 1$. This latter constraint would ensure that all vectors are in fact the same length. Our assumption only assures that they will be approximately the same length.

[4] Grossberg (1976) has addressed this problem in his very similar system. He has proved that if the patterns are sufficiently sparse, and/or when there are enough units in the cluster, then a system such as this will find a perfectly stable classification. He also points out that when these conditions don't hold, the classification can be unstable. Most of our work is with cases in which there is *no* perfectly stable classification and the number of patterns is *much* larger than the number of units in the inhibitory clusters.

Formal Analysis

Perhaps the simplest mathematical analysis that can be given of the competitive learning model under discussion involves the determination of the sets of *equilibrium states* of the system—that is, states in which the average inflow of weight to a particular line is equal to the average outflow of weight on that line. Let p_k be the probability that stimulus S_k is presented on any trial. Let v_{jk} be the probability that unit j wins when stimulus S_k is presented. Now we want to consider the case in which $\sum_k \Delta w_{ij} v_{jk} p_k = 0$, that is, the case in which the average change in the weights is zero. We refer to such states as *equilibrium states*. Thus, using the learning rule and averaging over stimulus patterns we can write

$$0 = g\sum_k \frac{c_{ik}}{n_k} p_k v_{jk} - g\sum_k w_{ij} p_k v_{jk}$$

which implies that at equilibrium

$$w_{ij}\sum_k p_k v_{jk} = \sum_k \frac{p_k c_{ik} v_{jk}}{n_k}$$

and thus

$$w_{ij} = \frac{\sum_k \frac{p_k c_{ik} v_{jk}}{n_k}}{\sum_k p_k v_{jk}}.$$

There are a number of important observations to note about this equation. First, note that $\sum_k p_k v_{jk}$ is simply the probability that unit j wins averaged over all stimulus patterns. Note further that $\sum_k p_k c_{ik} v_{jk}$ is the probability that input line i is active and unit j wins. Thus, the ratio $\frac{\sum_k p_k c_{ik} v_{jk}}{\sum_k p_k v_{jk}}$ is the conditional probability that line i is active given unit j wins, $p(\text{line}_i = 1 \mid \text{unit}_j \text{ wins})$. Thus, if all patterns are of the same size, i.e., $n_k = n$ for all k, then the weight w_{ij} becomes proportional to the probability that line i is active given unit j wins. That is,

$$w_{ij} \to \frac{1}{n} p(\text{line}_i = 1 \mid \text{unit}_j \text{ wins}).$$

We are now in a position to specify the response, at equilibrium, of unit j when stimulus S_l is presented. Let α_{jl} be the input to unit j in the face of stimulus S_l. This is simply the sum of weights on the active input lines. This can be written

$$\alpha_{jl} \to \sum_i w_{ij} c_{il} = \sum_i c_{il} \frac{\sum_k \frac{p_k c_{ik} v_{jk}}{n_k}}{\sum_k p_k v_{jk}}$$

which implies that at equilibrium

$$\alpha_{jl} = \frac{\sum_i p_i r_{li} v_{ji}}{\sum_i p_i v_{ji}}$$

where r_{li} represents the overlap between stimulus l and stimulus i, $r_{li} = \sum_k \frac{c_{kl} c_{ki}}{n_i}$. Thus, at equilibrium a unit responds most strongly to patterns that overlap other patterns to which the unit responds and responds most weakly to patterns that are far from patterns to which it responds. Finally, it should be noted that there is another set of restrictions on the value of v_{jk} — the probability that unit j responds to stimulus S_k. In fact, the competitive learning rule we have studied has the further restriction that

$$v_{jk} = \begin{cases} 1 & \alpha_{jk} > \alpha_{ik} \text{ for all } i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

Thus, in general, there are many solutions to the equilibrium equations described above. The competitive learning mechanisms can only reach those equilibrium states in which the above-stated relationships between the v_{jk} and the α_{jk} also hold.

Whenever the system is in a state in which, on average, the weights are not changing, we say that the system has reached an *equilibrium state*. In such a state the values of α_{jk} become relatively stable, and therefore, the values of v_{ik} become stable. When this happens, the system always responds the same way to a particular stimulus pattern. However, it is possible that the weights will be pushed out of

[5] There is a problem in that one can't be certain that the different clusters will discover different features. A slight modification of the system in which clusters "repel" one another can insure that different clusters find different features. We shall not pursue that further in this paper.

equilibrium by an unfortunate sequence of stimuli. In this case, the system can move toward a new equilibrium state (or possibly back to a previous one). Some equilibrium states are more stable than others in the sense that the v_{ik} become very unlikely to change values for long periods of time. In particular, this will happen whenever the largest α_{jk} is much larger than any other α_{ik} for all stimulus patterns S_k. In this case, small movements in the weight vector of one of the units is very unlikely to change which unit responds to which stimulus pattern. Such equilibrium states are said to be highly *stable*. We should expect, then, that after it has been learning for a period of time, the system will spend most of its time in the most highly stable of the equilibrium states. One good measure of the stability of an equilibrium state is given by the average amount by which the input to the winning units is greater than the response of all of the other units averaged over all patterns and all units in a cluster. This measure is given by T below:

$$T = \sum_k p_k \sum_{j,i} v_{jk} (\alpha_{jk} - \alpha_{ik}).$$

The larger the value of T, the more stable the system can be expected to be and the more time we can expect the system to spend in that state. Roughly, if we assume that the system moves into states which maximize T, we can show that this amounts to maximizing the overlap among patterns within a group while minimizing the overlap among patterns between groups. In the geometric analogy above, this will occur when the weight vectors point toward maximally compact stimulus regions that are as distant as possible from other such regions.

SOME EXPERIMENTAL RESULTS

Dipole Experiments

The essential structure that a competitive learning mechanism can discover is represented in the overlap of stimulus patterns. The simplest stimulus population in which stimulus patterns can overlap with one another is one constructed out of *dipoles*—stimulus patterns consisting of exactly two active elements and the rest inactive. If we have a total of N input units there are $N(N\text{-}1)/2$ possible dipole stimuli. Of course, if the actual stimulus population consists of all $N(N\text{-}1)/2$ possibilities, there is no structure to be discovered. There are no clusters for our units to point at (unless we have one unit for each of the possible stimuli, in which case we can point a weight vector at each of the

possible input stimuli). If, however, we restrict the possible dipole stimuli in certain ways, then there can be meaningful groupings of the stimulus patterns that the system can find. Consider, as an example, a case in which the stimulus lines could be thought of as forming a two-dimensional grid in which the only possible stimulus patterns were those which formed adjacent pairs in the grid. If we have an $N \times M$ grid, there are $N(M-1) + M(N-1)$ possible stimuli. Figure 4 shows one of the 24 possible adjacent dipole patterns defined on a 4×4 grid. We carried out a number of experiments employing stimulus sets of this kind. In most of these experiments we employed a two-layer system with a single inhibitory cluster of size two. Figure 5 illustrates the architecture of one of our experiments. The results of three runs with this architecture are illustrated in Figure 6, which shows the relative values of the weights for the two units. The values are shown laid out on a 4×4 grid so that weights are next to one another if the units with which they connect are next to one another. The relative values of the weights are indicated by the filling of the circles. If a circle is filled, that indicates that Unit 1 had the largest weight on that line. If the circle is unfilled, that means that Unit 2 had the largest weight on that line. The grids on the left indicate the initial configurations of the weights. The grids on the right indicate the final configurations of weights. The lines connecting the circles represent the possible stimuli. For example, the dipole stimulus pattern consisting of the upper left input line and the one immediately to the right of it is represented by the line connecting the upper-left circle in the grid with its right neighbor. The unit that wins when this stimulus is presented is indicated by the width of the line connecting the two circles. The wide line indicates

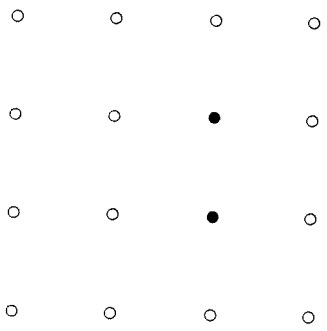

FIGURE 4. A dipole stimulus defined on a 4×4 matrix of input units. The rule for generating such stimuli is simply that any two adjacent units may be simultaneously active. Nonadjacent units may not be active and more than two units may not be simultaneously active. Active units are indicated by filled circles.

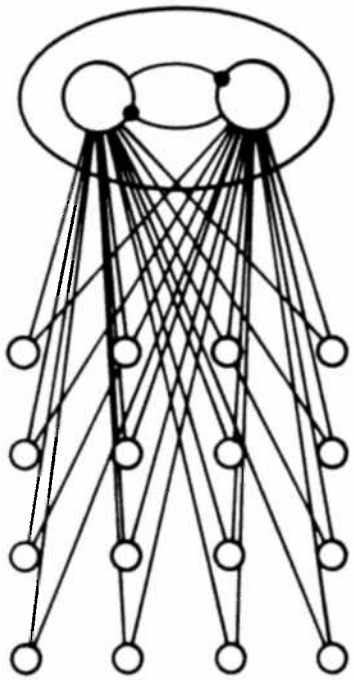

FIGURE 5. The architecture of a competitive learning system with 16 input units and one cluster of size two in the second layer.

that Unit 1 was the winner, the narrow line indicates that Unit 2 was the winner. It should be noted, therefore, that two unfilled circles must always be joined by a narrow line and two filled circles must always be joined by a wide line. The reason for this is that if a particular unit has more weight on both of the active lines then that unit *must* win the competition. The results clearly show that the weights move from a rather chaotic initial arrangement to an arrangement in which essentially all of those on one side of the grid are filled and all on the other side are unfilled. The border separating the two halves of the grid may be at any orientation, but most often it is oriented vertically and horizontally, as shown in the upper two examples. Only rarely is the orientation diagonal, as in the example in the lower right-hand grid. Thus, we have a case in which each unit has chosen a coherent half of the grid to which they respond. It is important to realize that as far as the competitive learning mechanism is concerned the sixteen input

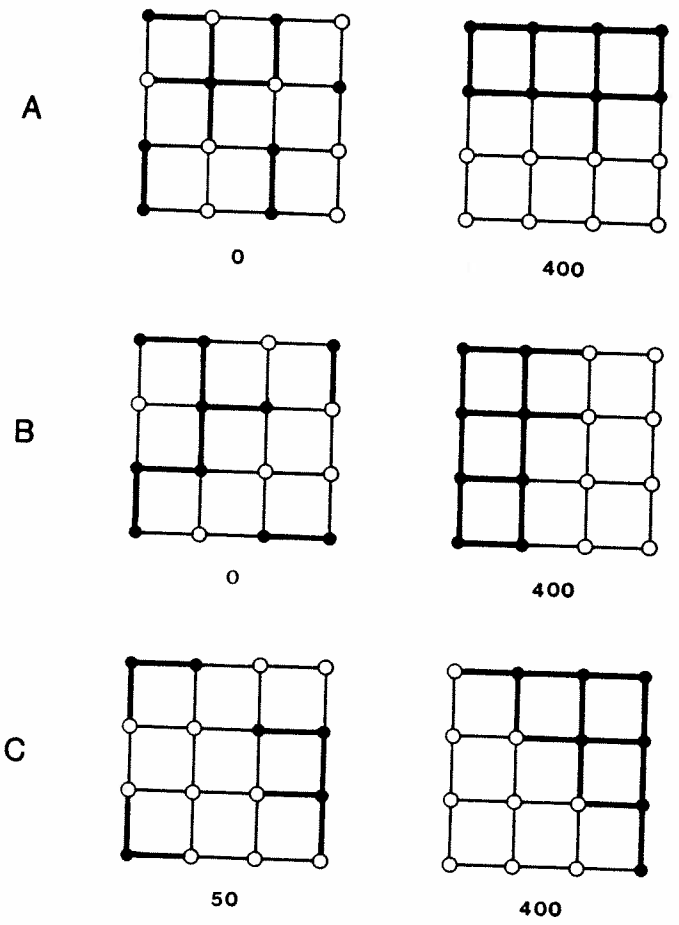

FIGURE 6. Relative weight values for the two members of the inhibitory cluster. *A*: The results for one run with the dipole stimuli defined over a two-dimensional grid. The left-hand grid shows the relative values of the weights initially and the right-hand grid shows the relative values of the weights after 400 trials. A filled circle means that Unit 1 had the larger weight on the corresponding input. An unfilled circle means that Unit 2 had the larger weight. A heavy line connecting two circles means that Unit 1 responded to the stimulus pattern consisting of the activation of the two circles, and a light line means that Unit 2 won the corresponding pattern. In this case the system has divided the grid horizontally. *B*: The results for a second run under the same conditions. In this case the system has divided the grid horizontally. *C*: The results for a third run. In this case the left-hand grid represents the state of the system after 50 trials. Here the grid was divided diagonally.

lines are unordered. The two-dimensional grid-like arrangement exists only in the statistics of the population of stimulus patterns. Thus, the system has *discovered* the dimensional structure inherent in the stimulus population and has devised binary feature detectors that tell which half of the grid contains the stimulus pattern. Note, each unit responds to roughly half of the stimulus patterns. Note also that while some units break the grid vertically, some break the grid horizontally, and some break it diagonally; a combination of several clusters offers a rather more precise classification of a stimulus pattern.

In other experiments, we tried clusters of other sizes. For example, Figure 7 shows the results for a cluster of size four. It shows the initial configuration and its sequence of evolution after 100, 200, 400, 800, and after 4000 training trials. Again, initially the regions are chaotic. After training, however, the system settles into a state in which stimuli in compact regions of the grid are responded to by the same units. It can be seen, in this case, that the trend is toward a given unit responding to a maximally compact group of stimuli. In this experiment, three of the units settled on compact square regions while the remaining one settled on two unconnected stimulus regions. It can be shown that the state into which the system settled does not quite maximize the value T, but does represent a relatively stable equilibrium state.

In the examples discussed thus far, the system, to a first approximation, settled on a highly compact representation of the input patterns in which all patterns in a region are captured by one of the units. The grids discussed above have all been two-dimensional. There is no need to restrict the analysis to a two-dimensional grid. In fact, a two-unit cluster will, essentially, pass a plane through a space of any dimensionality. There is a preference for planes perpendicular to the axes of the spaces. Figure 8 shows a typical result for the system learning a three-dimensional space. In the case of three dimensions, there are three equally good planes which can be passed through the space and, depending on the starting directions of the weight vectors and on the sequence of stimuli, different clusters will choose different ones of these planes. Thus, a system which receives input from a set of such clusters will be given information as to which *quadrant* of the space in which the pattern appears. It is important to emphasize that the coherence of the space is *entirely* in the choice of input stimuli, *not* in the architecture of the competitive learning mechanism. The system *discovers* the spatial structure in the input lines.

Formal analysis. For the dipole examples described above, it is possible to develop a rather precise characterization of the behavior of the competitive learning system. Recall our argument that the most stable equilibrium state (and therefore the one the system is *most* likely to

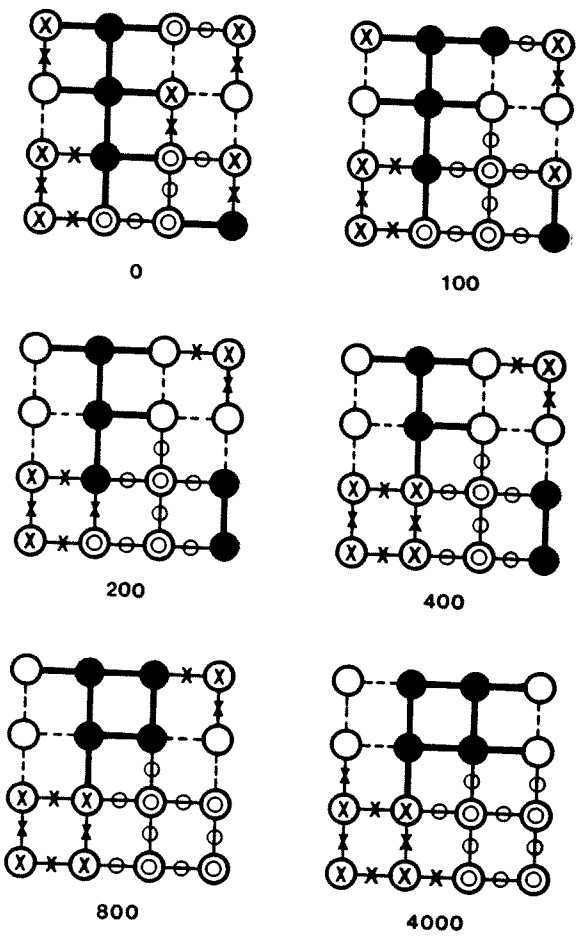

FIGURE 7. The relative weights of each of the four elements of the cluster after 0, 100, 200, 400, 800, and 4000 stimulus presentations.

end up in) is the one that maximizes the function

$$T = \sum_k p_k \sum_{j,i} v_{jk} (\alpha_{jk} - \alpha_{ik}).$$

Now, in the dipole examples, all stimulus patterns of the stimulus population are equally likely (i.e., $p_k = 1/N$), all stimulus patterns involve two active lines, and for every stimulus pattern in the population of patterns there are a fixed number of other stimulus patterns in

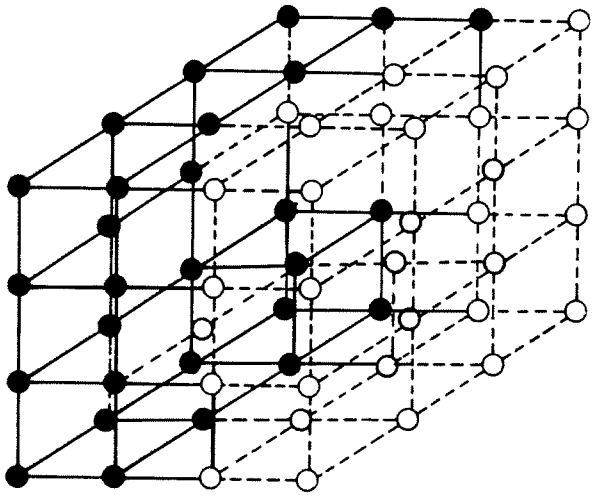

FIGURE 8. The relative weights for a system in which the stimulus patterns were chosen from a three-dimensional grid after 4000 presentations.

the population which overlap it.[6] This implies that $\sum_k^N r_{kj} = R$ for all j. With these assumptions, it is possible to show that maximizing T is equivalent to minimizing the function

$$\sum_i^M \frac{B_i}{N_i}$$

(see appendix for derivation), where N_i is the number of patterns on which unit i wins, M is the number of units in the cluster, and B_i is the number of cases in which unit i responds to a particular pattern and does not respond to a pattern which overlaps it. This is the number of *border* patterns to which unit i responds. Formally, we have

$$B_i = \sum_j^N \sum_k^N v_{ij}(1-v_{ik}) \text{ for } r_{jk} > 0.$$

From this analysis, it is clear that the most stable states are ones in which the size of the border is minimized. Since total border region is minimized when regions are spherical, we can conclude that in a situation in which stimulus pairs are drawn from adjacent points in a

[6] Note that this latter condition does not quite hold for the examples presented above due to edge effects. It is possible to eliminate edge effects by the use of a torus. We have carried out experiments on tori as well, and the results are essentially the same.

high-dimensional hyperspace, our competitive learning mechanism will form essentially spherical regions that partition the space into one such spherical region for each element of the cluster.

Another result of our simulations which can be explained by these equations is the tendency for each element of the cluster to capture roughly equally sized regions. This results from the interconnectedness of the stimulus population. The result is easiest in the case in which $M = 2$. In this case, the function we want to minimize is given by

$$\frac{B_1}{N_1} + \frac{B_2}{N_2}.$$

Now, in the case of $M = 2$, we have $B_1 = B_2$, since the two regions must border on one another. Moreover, we have $N_1 + N_2 = N$, since every pattern is either responded to by Unit 1 or Unit 2. Thus, we want to minimize the function

$$B\left(\frac{1}{N_1} + \frac{1}{N-N_1}\right).$$

This function is minimized when $N_1 = N/2$. Thus, there are two pressures which determine the performance of the system in these cases:

- There is a pressure to reduce the number of border stimuli to a minimum.
- There is a pressure to divide the stimulus patterns among the units in a way that depends on the total amount of weight that unit has. If two units have the same amount of weight, they will capture roughly equal numbers of equally likely stimulus patterns.

Learning Words and Letters

It is common practice to handcraft networks to carry out particular tasks. Whenever one creates such a network that performs a task rather successfully, the question arises as to how such a network might have evolved. The word perception model developed in McClelland and Rumelhart (1981) and Rumelhart and McClelland (1982) is one such case-in-point. That model offers rather detailed accounts of a variety of word perception experiments, but it was crafted to do its job.

How could it have evolved naturally? Could a competitive learning mechanism create such a network?

Let's begin with the fact that the word perception model required a set of position-specific letter detectors. Suppose that a competitive learning mechanism is faced with a set of words—to what features would the system learn to respond? Would it create position-specific letter detectors or their equivalent? We proceeded to answer this question by again viewing the lower level units as forming a two-dimensional grid. Letters and words could then be presented by activating those units on the grid corresponding to the points of a standard CRT font. Figure 9 gives examples of some of the stimuli used in our experiments. The grid we used was a 7×14 grid. Each letter occurred in a 7×5 rectangular region on the grid. There was room for two letters with some space in between, as shown in the figure. We then carried out a series of experiments in which we presented a set of word and/or letter stimuli to the system allowing it to extract relevant features.

Before proceeding with a description of our experiments, it should be mentioned that these experiments required a slight addition to the competitive learning mechanism. The problem was that, unlike the dipole stimuli, the letter stimuli only sparsely covered the grid and many of the units in the lower level never became active at all. Therefore, there was a possibility that, by chance, one of the units would have most of its weight on input lines that were never active, whereas another unit may have had most of its weight on lines common to all of the stimulus patterns. Since a unit never learns unless it wins, it is

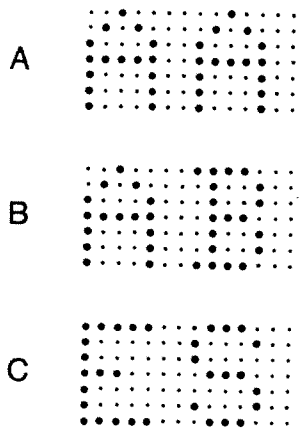

FIGURE 9. Example stimuli for the word and letter experiments.

possible that one of the units will never win, and therefore never learn. This, of course, takes the competition out of competitive learning. This situation is analogous to the situation in the geometric analogy in which all of the stimulus points are relatively close together on the hypersphere, and one of the weight vectors, by chance, points near the cluster while the other one points far from the stimuli. (See Figure 10). It is clear that the more distant vector is not closest to any stimulus and thus can never move toward the collection. We have investigated two modifications to the system which deal with the problem. One, which we call the leaky learning model, modifies the learning rule to state that *both* the winning *and* the losing units move toward the presented stimulus: the close vector simply moves much further. In symbols this suggests that

$$\Delta w_{ij} = \begin{cases} g_l \dfrac{c_{ik}}{n_k} - g_l w_{ij} & \text{if unit } j \text{ loses on stimulus } k \\ g_w \dfrac{c_{ik}}{n_k} - g_w w_{ij} & \text{if unit } j \text{ wins on stimulus } k \end{cases}$$

where g_l is the learning rate for the losing units, g_w is the learning rate for the winning unit, and where $g_l << g_w$. In our experiments we made

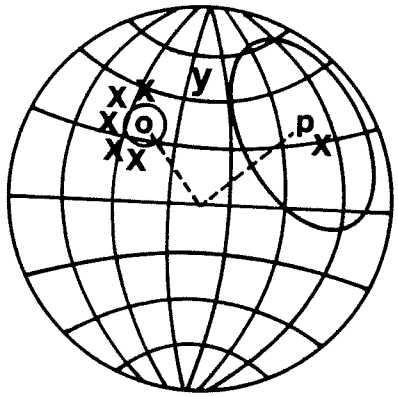

FIGURE 10. A geometric interpretation of changes in stimulus sensitivity. The larger the circle around the head of the weight vector the more sensitive the unit. The decision as to which unit wins is made on the basis of the distance from the circle rather than from the head of the weight vector. In the example, the stimulus pattern indicated by the y is actually closer to the head of one vector $\mathbf{o}$, but since it is closer to the circle surrounding vector $\mathbf{p}$, unit p would win the competition.

g_l an order of magnitude smaller than g_w. This change has the property that it slowly moves the losing units into the region where the actual stimuli lie, at which point they begin to capture some units and the ordinary dynamics of competitive learning take over.

The second method is similar to that employed by Bienenstock, Cooper, and Munro (1982), in which a unit modulates its own sensitivity so that when it is not receiving enough inputs, it becomes increasingly sensitive. When it is receiving too many inputs, it decreases its sensitivity. This mechanism can be implemented in the present context by assuming that there is a threshold and that the relevant activation is the degree to which the unit exceeds its threshold. If, whenever a unit fails to win it decreases its threshold and whenever it does win it increases its threshold, then this method will also make all of the units eventually respond, thereby engaging the mechanism of competitive learning. This second method can be understood in terms of the geometric analogy that the weight vectors have a circle surrounding the end of the vector. The relevant measure is not the distance to the vector itself but the distance to the circle surrounding the vector. Every time a unit loses, it increases the radius of the circle; every time it wins, it decreases the radius of the circle. Eventually, the circle on the losing unit will be large enough to be closer to some stimulus pattern than the other units.

We have used both of these mechanisms in our experiments and they appear to result in essentially similar behavior. The former, the leaky learning method, does not alter the formal analysis as long as the ratio g_l/g_w is sufficiently small. The varying threshold method is more difficult to analyze and may, under some circumstances, distort the competitive learning process somewhat. After this diversion, we can now return to our experiments on the development of word/position-specific letter detectors and other feature detectors.

Position-specific letter detectors. In our first experiment, we presented letter pairs drawn from the set: *AA, AB, BA,* and *BB*. We began with clusters of size two. The results were unequivocal. The system developed position-specific letter detectors. In some experimental runs, one of the units responded whenever *AA* or *AB* was presented, and the other responded whenever *BA* or *BB* was presented. In this case, Unit 1 represents an *A* detector in position 1 and Unit 2 represents a *B* detector for position 1. Moreover, as in the word perception model, the letter detectors are, of course, in a mutually inhibitory pool. On other experimental runs, the pattern was reversed. One of the units responded whenever there was an *A* in the second position and the other unit responded whenever there was a *B* in the second position. Figure 11 shows the final configuration of weights for one of

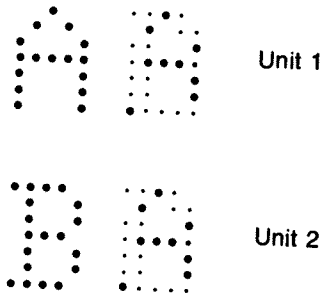

FIGURE 11. The final configuration of weights for a system trained on the stimulus patterns *A, B, C, D*.

our experimental runs. Note that although the units illustrated here respond *only* to the letter in the first position, there is still weight on the active lines in the second position. It is just that the weights on the first position differentiate between *A* and *B*, whereas those on the second position respond equally to the two letters. In particular, as suggested by our formal analysis, asymptotically the weights on a given line are proportional to the probability that that line is active when the unit wins. That is, $w_{ij} \rightarrow p(\text{unit}_i = 1 \mid \text{unit}_j \text{ wins})$. Since the lower level units unique to *A* occur equally as often as those unique to *B*, the weights on those lines are roughly equal. The input lines common to the two letters are on twice as often as those unique to either letter, and hence, they have twice as much weight. Those lines that never come on reach zero weight.

Word detection units. In another experiment, we presented the same stimulus patterns, but increased the elements in the cluster from two to four. In this case, each of the four level-two units came to respond to one of the four input patterns—in short, the system developed *word detectors*. Thus, if layer two were to consist of a number of clusters of various sizes, large clusters with approximately one unit per word pattern will develop into word detectors, while smaller clusters with approximately the number of letters per spatial position will develop into position-specific letter detectors. As we shall see below, if the number of elements of a cluster is substantially less than the number of letters per position, then the cluster will come to detect position-specific letter features.

Effects of number of elements per serial position. In another experiment, we varied the number of elements in a cluster and the number of letters per serial position. We presented stimulus patterns drawn

from the set: *AA, AB, AC, AD, BA, BB, BC, BD.* In this case, we found that with clusters of size two, one unit responded to the patterns beginning with *A* and the other responded to those beginning with *B*. In our previous experiment, when we had the same number of letters in each position, we found that the clusters were indifferent as to which serial position they responded. Some responded to position 1 and others to position 2. In this experiment, we found that a two-element cluster always becomes a letter detector specific to serial position in which two letters vary. Similarly, in the case of clusters of size four we found that they always became letter detectors for the position in which four letters varied. Thus, in this case one responded to an *A* in the second position, one responded to a *B* in the second position, one responded to a *C* in the second position, and one responded to a *D* in the second position. Clearly, there are two natural ways to cluster the stimulus patterns—two levels of structure. If the patterns are to be put in two categories, then the binary feature *A* or *B* in the first position is the relevant distinction. On the other hand, if the stimuli are to be grouped into four groups, the four value feature determining the second letter is the relevant distinction. The competitive learning algorithm can discover either of the levels of structure—depending on the number of elements in a cluster.

Letter similarity effects. In another experiment, we studied the effects of letter similarity to look for units that detect letter features. We presented letter patterns consisting of a letter in the first position only. We chose the patterns so they formed two natural clusters based on the similarity of the letters to one another. We presented the letters *A, B, S,* and *E*. The letters were chosen so that they fell naturally into two classes. In our font, the letters *A* and *E* are quite similar and the letters *B* and *S* are very similar. We used a cluster of size two. Naturally, one of the units responded to the *A* or the *E* while the other unit responded to the *B* or the *S*. The weights were largest on those features of the stimulus pairs which were common among each of these similar pairs. Thus, the system developed subletter-size feature detectors for the features relevant to the discrimination.

Correlated teaching inputs. We carried out one other set of experiments with the word/letter patterns. In this case, we used clusters of size two and presented stimuli drawn from the set: *AA, BA, SB, EB.* Note that on the left-hand side, we have the same four letters as we had in the previous experiment, but on the right-hand side we have only two patterns; these two patterns are correlated with the letter in the first position. An *A* in the second position means that the first position contains either an *A* or a *B*, whereas a *B* in the second position

means that the first position contains either an *S* or an *E*. Note further that those correlations between the first and second positions are in opposition to the "natural" similarity of the letters in the first serial position. In this experiment, we first trained the system on the four stimuli described above. Since the second serial position had only two letters in it, the size-two cluster became a position-specific letter detector for the second serial position. One unit responded to the *A* and one to the *B* in the second position. Notice that the units are also responding to the letters in the first serial position as well. One unit is responding to an *A* or a *B* in the first position while the other responds to an *E* or an *S*. Figure 12 shows the patterns of weights developed by the two units. After training, the system was then presented patterns containing only the first letter of the pair and, as expected, the system had learned the "unnatural" classification of the letters in the first position. Here the strong correlation between the first and second position led the competitive learning mechanism to override the strong correlation between the highly similar stimulus patterns in the first serial position. This suggests that even though the competitive learning system is an "unsupervised" learning mechanism, one can control what it learns by controlling the statistical structure of the stimulus patterns being presented to it. In this sense, we can think of the right-hand letter in this experiment as being a kind of *teaching* stimulus aimed at determining the classification learned for other aspects of the stimulus. It should also be noted that this teaching mechanism is essentially the same as the so-called errorless learning procedure used by Terrace (1963) in training pigeons to peck a certain color key by associating that color with a response situation where their pecking is determined by other factors. As we shall see below, this correlational teaching mechanism is useful in allowing the competitive learning mechanism to discover features which it otherwise would be unable to discover.

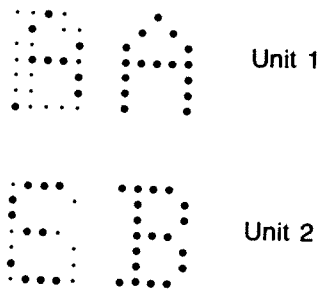

FIGURE 12. The pattern of weights developed in the correlated learning experiment.

Horizontal and Vertical Lines

One of the classically difficult problems for a linear threshold device like a perceptron is to distinguish between horizontal and vertical lines. In general, horizontal and vertical lines are not linearly separable and require a multilayer perceptron system to distinguish them. One of the goals of the competitive learning device is for it to discover features that, at a higher level of analysis, might be useful for discriminating patterns which might not otherwise be discriminable with a linear threshold-type device. It is therefore of some interest to see what kinds of features the competitive learning mechanism discovers when presented with a set of vertical and horizontal lines. In the following discussion, we chronicle a series of experiments on this problem. Several of the experiments ended in failure, but we were able to discover a way in which competitive learning systems can be put together to build a hierarchical feature detection system capable of discriminating vertical and horizontal lines. We proceed by sketching several of our failures as well as our successes because the way in which the system fails is elucidating. It should be noted at the outset that our goal is not so much to present a model of how the human learns to distinguish between vertical and horizontal lines (indeed, such a distinction is probably prewired in the human system), but rather to show how competitive learning can discover features which allow for the system to learn distinctions with multiple layers of units that cannot be learned by single-layered systems. Learning to distinguish vertical and horizontal lines is simply a paradigm case.

In this set of experiments, we represented the lower level of units as if they were on a 6×6 grid. We then had a total of 12 stimulus patterns, each consisting of turning on six Level 1 units in a row on the grid. Figure 13 illustrates the grid and several of the stimulus patterns. Ideally, one might hope that one of the units would respond whenever a vertical line is presented; the other would respond whenever a horizontal line is presented. Unfortunately, a little thought indicates that this is impossible. Since every input unit participates in exactly one vertical and one horizontal line, there is no configuration of weights which will distinguish vertical from horizontal. This is exactly why no linear threshold device can distinguish between vertical and horizontal lines in one level. Since that must fail, we might hope that some clusters in the competitive learning device will respond to vertical lines by assigning weights as illustrated in Figure 14. In this case, one unit of the pair would respond whenever the first, second, or fourth vertical line was presented, and another would respond whenever the third, fifth, or sixth vertical line was presented; since both units would

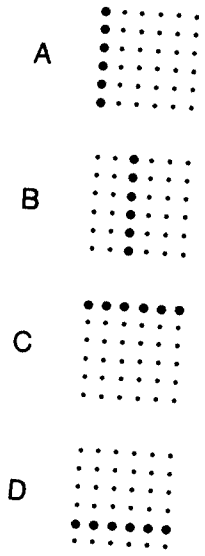

FIGURE 13. Stimulus patterns for the horizontal/vertical discrimination experiments.

receive about the same input in the face of a horizontal line, we might expect that sometimes one and sometimes the other would win the competition but that the primary response would be to vertical lines. If other clusters settled down similarly to horizontal lines, then a unit at the third level looking at the output of the various clusters could distinguish vertical and horizontal. Unfortunately, that is not the pattern of weights discovered by the competitive learning mechanism. Rather, a typical pattern of weights is illustrated in Figure 15. In this arrangement, each cluster responds to exactly three horizontal and three vertical lines. Such a cluster has lost all information that might distinguish vertical from horizontal. We have discovered a feature of absolutely no use in this distinction. In fact, such features systematically throw away the information relevant to horizontal vs. vertical. Some further thought indicates why such a result occurred. Note, in particular, that two horizontal lines have exactly *nothing* in common. The grid that we show in the diagrams is merely for our convenience. As far as the units are concerned there are 36 unordered input units; sometimes some of those units are active. Pattern similarity is determined entirely by pattern overlap. Since horizontal lines don't intersect, they have no units in common, thus they are not seen as similar at all. However, every horizontal line intersects with every vertical line and thus has much more in common with vertical lines than with other horizontal

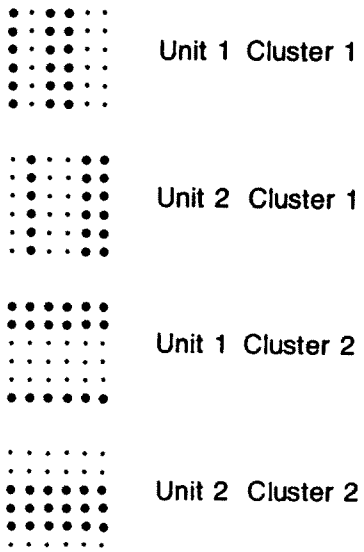

FIGURE 14. A possible weight configuration which could distinguish vertical from horizontal.

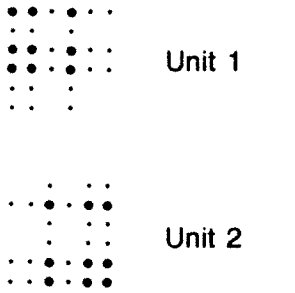

FIGURE 15. A typical configuration of weights for the vertical/horizontal discrimination.

ones. It is this similarity that the competitive learning mechanism has discovered.

Now, suppose that we change the system somewhat. Suppose that we "teach" the system the difference between vertical and horizontal (as we did in the previous experiments with letter strings). In this experiment we used a 12×6 grid. On the right-hand side of the grid we presented either a vertical or a horizontal line, as we did before. On

the left-hand side of the grid we always presented the uppermost horizontal line whenever any horizontal line was presented on the right-hand grid, and we always presented the vertical line furthest to the left on the left-hand grid whenever we presented any vertical line on the right-hand side of the grid. We then had a cluster of two units receiving inputs from all $12 \times 6 = 72$ lower level units. (Figure 16 shows several of the stimulus patterns.)

As expected, the two units soon learned to discriminate between vertical and horizontal lines. One of the units responded whenever a vertical line was presented and the other responded whenever a horizontal line was presented. They were responding, however, to the pattern on the left-hand side rather than to the vertical and horizontal pattern on the right. This too should be expected. Recall that the value of the w_{ij} approaches a value which is proportional to the probability that input unit i is active, given that unit j won the competition. Now, in the case of the unit that responds to vertical lines for example, every unit on the right-hand grid occurs equally often so that all of the weights connecting to units in that grid have equal weights. The same is true for the unit responding to the horizontal line. The weights on

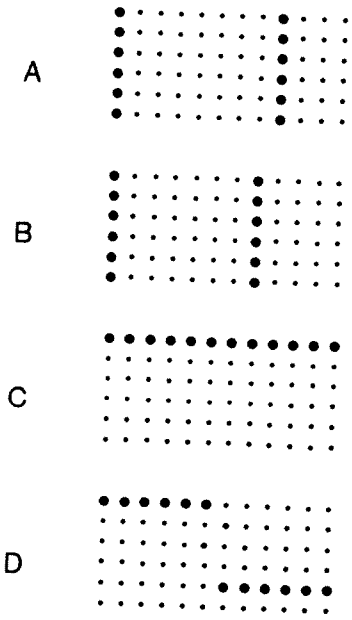

FIGURE 16. Stimulus patterns for the vertical/horizontal discrimination experiments with a correlated "teaching" input on the right-hand side.

the right-hand grid are identical for the two cluster members. Thus, when the "teacher" is turned off, and only the right-hand figure is presented, the two units respond randomly and show no evidence of having learned the horizontal/vertical distinction.

Suppose, however, that we have four, rather than two, units in the level-two clusters. We ran this experiment and found that of the four units, two of them divided up the vertical patterns and two of them divided up the horizontal patterns. Figure 17 illustrates the weight values for one of our runs. One of the units took three of the vertical line patterns; another unit took three other vertical patterns. A third unit responded to three of the horizontal line patterns, and the last unit responded to the remaining three horizontal lines. Moreover, after we took away the "teaching" pattern, the system continued to classify the vertical and horizontal lines just as it did when the left-hand "teaching" pattern was present.

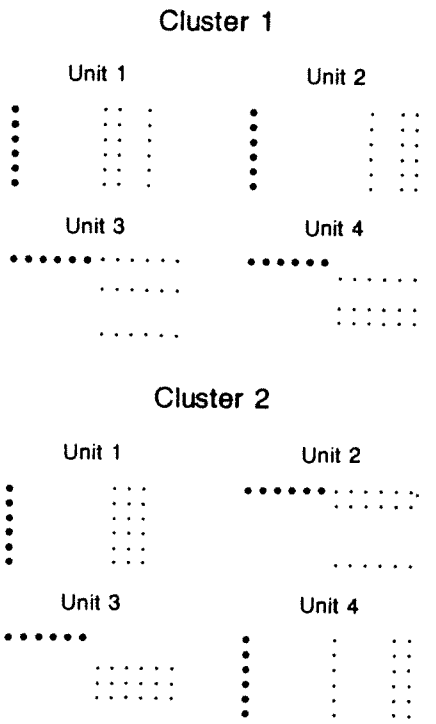

FIGURE 17. The weight values for the two clusters of size four for the vertical/horizontal discrimination experiment with a correlated "teaching" stimulus.

In one final experiment with vertical and horizontal lines, we developed a three-level system in which we used the same stimulus patterns as in the previous experiment; the only difference was that we had *two* clusters of four units at the second level and one cluster of two units at the third level. Figure 18 shows the architecture employed. In this case, the two four-element clusters each learned to respond to subsets of the vertical and horizontal lines as in the previous experiment. The two clusters generally responded to different subsets, however. Thus, when the upper horizontal line was presented, Unit 1 of the first cluster responded and Unit 3 of the second cluster responded. When the bottom horizontal line was presented, Unit 1 of the first cluster responded again, but Unit 4 of the second cluster also responded. Thus, the cluster of size two at the highest level was receiving a kind of dipole stimulus. It has four inputs and on any trial, two of them are active. As with our analysis of dipole stimuli, we know that stimuli that overlap are always put in the same category. Note that when a vertical line is presented, one of the two units in each of the middle layers of clusters that responds to vertical lines will become active, and that none of the units that respond to horizontal lines will ever be active; thus, this means that there are two units in each middle layer cluster that respond to vertical lines. Whenever a vertical line is presented, one of the units in each cluster will become active. None of the horizontal units will ever be active in the face of a vertical stimulus. Thus, one of the units at the highest level learns to respond whenever a vertical line is presented, and the other unit responds whenever a horizontal line is

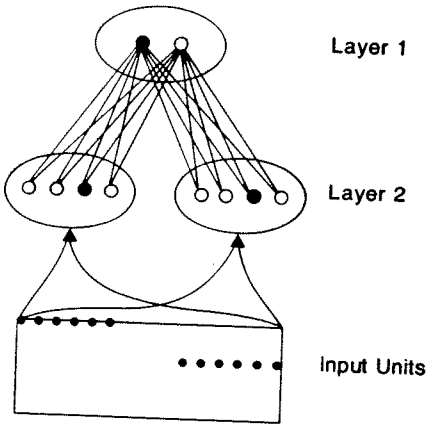

FIGURE 18. The architecture for the three-level horizontal/vertical discrimination experiment.

presented. Once the system has been trained, this occurs despite the absence of the "teaching" stimulus. Thus, what we have shown is that the competitive learning mechanism can, under certain conditions, develop feature detectors which allow the system to distinguish among patterns that are not differentiable by a simple linear unit in one level.

CONCLUSION

We have shown how a very simple competitive mechanism can discover a set of feature detectors that capture important aspects of the set of stimulus input patterns. We have also shown how these feature detectors can form the basis of a multilayer system that can serve to learn categorizations of stimulus sets that are not linearly separable. We have shown how the use of correlated stimuli can serve as a kind of "teaching" input to the system to allow the development of feature detectors which would not develop otherwise. Although we find the competitive learning mechanism a very interesting and powerful learning principle, we do not, of course, imagine that it is the only learning principle. Competitive learning is an essentially nonassociative, statistical learning scheme. We certainly imagine that other kinds of learning mechanisms will be involved in the building of associations among patterns of activation in a more complete neural network. We offer this analysis of these competitive learning mechanisms to further our understanding of how simple adaptive networks can discover features important in the description of the stimulus environment in which the system finds itself.

ACKNOWLEDGMENTS

This research was supported by grants from the System Development Foundation and by Contract N00014-79-C-0323, NR 667-437 with the Personnel and Training Research Programs of the Office of Naval Research.

APPENDIX

For the case of homogeneous *dipole* stimulus patterns, it is possible to derive an expression for the most *stable* equilibrium state of the system. We say that a set of dipole stimulus patterns is homogeneous if (a) they are equally likely and (b) for every input pattern in the set there are a fixed number of other input patterns that overlap them. These conditions were met in our simulations. Our measure of stability is given by

$$T = \sum_k p_k \sum_j \sum_i v_{jk} (\alpha_{jk} - \alpha_{ik}).$$

Since $p_k = \frac{1}{N}$, we can write

$$T = \frac{1}{N} \sum_i \sum_j \sum_k v_{jk} \alpha_{jk} - \frac{1}{N} \sum_i \sum_j \sum_k v_{jk} \alpha_{ik}.$$

Summing the first portion of the equation over i and the second over j we have

$$T = \frac{M}{N} \sum_j \sum_k v_{jk} \alpha_{jk} - \frac{1}{N} \sum_l \sum_k \alpha_{ik} \sum_j v_{jk}.$$

Now note that when $p_k = 1/N$, we have $\alpha_{ik} = \sum_j r_{kj} v_{ij} / \sum_l v_{kl}$. Furthermore, $\sum_l v_{lk} = 1$ and $\sum_k v_{lk} = N_l$, where N_l is the number of patterns captured by unit l. Thus, we have

$$T = \frac{M}{N} \sum_j \sum_k v_{jk} \alpha_{jk} - \frac{1}{N} \sum_i \sum_k \frac{\sum_l r_{kl} v_{il}}{N_l}.$$

Now, since all stimuli are the same size, we have $r_{ij} = r_{ji}$. Moreover, since all stimuli have the same number of neighbors, we have $\sum_i r_{ij} = \sum_j r_{ij} = R$, where R is a constant determined by the dimensionality of the stimulus space from which the dipole stimuli are drawn. Thus, we have

$$T = \frac{M}{N} \sum_j \sum_k v_{jk} \alpha_{jk} - \frac{R}{N} \sum_i \frac{\sum_l v_{il}}{N_l},$$

and we have

$$T = \frac{M}{N}\sum_j\sum_k v_{jk}\alpha_{jk} - \frac{RM}{N}.$$

Since R, M, and N are constants, we have that T is maximum whenever $T' = \sum_j\sum_k v_{jk}\alpha_{jk}$ is maximum. Now substituting for α_{jk}, we can write

$$T' = \sum_j \frac{1}{N_j}\sum_k\sum_l r_{kl}\, v_{jk}\, v_{jl}.$$

We can now substitute for the product $v_{jk}\, v_{jl}$ the term $v_{jk} - v_{jk}(1 - v_{jl})$. We then can write

$$T' = \sum_j \frac{1}{N_j}\sum_k\sum_l r_{kl}\, v_{jk} - \sum_j \frac{1}{N_j}\sum_k\sum_l r_{kl}\, v_{jk}(1 - v_{jl}).$$

Summing the first term of the equation first over l, then over k, and then over j, gives us

$$T' = MR - \sum_j \frac{1}{N_j}\sum_k\sum_l r_{kl}\, v_{jk}(1 - v_{jl}).$$

Now, recall that r_{kl} is given by the degree of stimulus overlap between stimulus l and stimulus k. In the case of dipoles there are only three possible values of r_{kl}:

$$r_{kl} = \begin{cases} 0 & \text{no overlap} \\ 1 & k=l \\ 1/2 & \text{otherwise} \end{cases}$$

Now, the second term of the equation for T' is 0 if either $r_{kl} = 0$ or if $v_{jk}(1 - v_{jl}) = 0$. Since v_{ik} is either 1 or 0, this will be zero whenever $j=l$. Thus, for all nonzero cases in the second term we have $r_{kl} = \frac{1}{2}$. Thus we have

$$T' = MR - \frac{1}{2}\sum_j \frac{1}{N_j}\sum_k\sum_l v_{jk}(1 - v_{jl}).$$

Finally, note that $\sum_k\sum_l v_{jk}(1 - v_{jl})$ is 1 and r_{kl} is $\frac{1}{2}$ in each case in which different units capture neighboring patterns. We refer to this as a case of *bad neighbors* and let B_j designate the number of bad neighbors for unit j. Thus, we have

$$T' = MR - \frac{1}{2}\sum_j \frac{B_j}{N_j}.$$

Finally, we can see that T' will be a maximum whenever $T'' = \sum_j \frac{B_j}{N_j}$ is minimum. Thus, minimizing T'' leads to the maximally stable solution in this case.