Human Category Learning: Implications for Backpropagation Models

Kruschke, J. K. (1993). Human category learning: implications for backpropagation models. *Connection Science*, **5**, 3-36.

JOHN K. KRUSCHKE

(Received for publication 10 June 1992; revised paper accepted 29 October 1992)

Backpropagation (Rumelhart et al., *1986a) was proposed as a general learning algorithm for multi-layer perceptrons. This article demonstrates that a standard version of backprop fails to attend selectively to input dimensions in the same way as humans, suffers catastrophic forgetting of previously learned associations when novel exemplars are trained, and can be overly sensitive to linear category boundaries. Another connectionist model, ALCOVE (Kruschke 1990, 1992), does not suffer those failures. Previous researchers identified these problems; the present article reports quantitative fits of the models to new human learning data. ALCOVE can be functionally approximated by a network that uses linear-sigmoid hidden nodes, like standard backprop. It is argued that models of human category learning should incorporate quasi-local representations and dimensional attention learning, as well as error-driven learning, to address simultaneously all three phenomena.*

KEYWORDS: Backpropagation, catastrophic forgetting, categorization, coarse coding, condensation, dimensional attention, error-driven learning, filtration, linear boundaries, local representation.

1. Introduction

Standard backpropagation (Rumelhart *et al.*, 1986a), or 'backprop', was originally proposed as a learning mechanism for multi-layer perceptrons (Rosenblatt, 1958; Minsky & Papert, 1969). Its main goal was to learn internal representations that could mediate complex mappings between inputs and outputs, as evidenced by the very title of the landmark report of Rumelhart *et al.* (1986a): 'Learning internal representations . . . '. Many papers have been written devoted to the analysis of the internal representations discovered by backprop (e.g. Elman, 1989; Hanson & Burr, 1991; Rosenberg, 1987).

Rumelhart *et al.* (1986a) did not address the question of whether backprop could model the course of human learning. This question can be asked at the neural or molar levels. It is generally (though not universally) agreed that backprop cannot be *directly* implemented in real neurons, given present-day knowledge of neural function (e.g. Grossberg, 1987; Rumelhart *et al.*, 1986b, p. 536; Stork,

J. K. Kruschke, Department of Psychology and Cognitive Science Program, Indiana University, Bloomington, IN 47405-4201, USA. E-mail: kruschke@ucs.indiana. edu.

1989). The general sentiment of most users of backprop was clearly expressed by Lehky and Sejnowski (1988, p. 454): "No biological significance is claimed for the algorithm (back propagation) by which the network developed, but, rather, the focus of interest is on the resulting mature network." Ultimately, neural plausibility is desirable for any model of behavior, especially for network models that are ostensibly 'brain style' and 'neurally inspired' (Rumelhart & McClelland, 1986). Nevertheless, there is a long history of learning models that make no attempt to contact neural functioning, although there is an implicit recognition that neural mechanisms must somehow implement them (e.g. Bower & Hilgard, 1981). The goal of such models is to capture accurately some of the molar learning behavior observed in people. Therefore, despite the neural implausibility of backprop, we can ask whether it reflects the course of learning at the molar, behavioral level.

Since 1986 many researchers have used backprop in models of human learning at the molar level. Several reports have emphasized its success (e.g. Cohen *et al.*, 1990; McClelland & Jenkins, 1991; Seidenberg & McClelland, 1989; Sejnowski & Rosenberg, 1988; Taraban *et al.*, 1989), and others have emphasized its failures (e.g. Gluck, 1991; McCloskey & Cohen, 1989; Pavel *et al.*, 1989; Ratcliff, 1990). This article isolates three failures of standard backprop to model human category learning. By 'category learning' I mean situations in which people learn to associate category labels with stimuli. First, backprop fails to learn category distinctions for which only a few stimulus dimensions are relevant faster than distinctions for which a large number of stimulus dimensions are relevant. In other words, backprop fails to attend selectively to stimulus dimensions the same way people do. Second, backprop suffers 'catastrophic forgetting' of previously learned associations when new associations are trained. Third, standard backprop learns linearly separable categorizations faster than non-linearly separable ones in some situations where people do not.

An alternative model, called ALCOVE (Kruschke, 1990, 1992), overcomes these problems. The model was motivated by a molar-level psychological theory, Nosofsky's (1986) generalized context model (GCM), rather than by neuron-like perceptrons. ALCOVE is closely related to the structure of standard backprop, in that it is also a feed-forward network that learns using gradient descent on error, but unlike backprop it has explicit attention strengths on the input dimensions, and it uses hidden nodes with a different activation function than used in backprop.

I would like to emphasize from the outset that the cause of backprop's problems is not the error-driven learning mechanism, but its particular architecture. One goal of this article is to demonstrate that these two models, though similar, generate very different behavior. I also show that backprop can be modified to mimic the functionality of ALCOVE, and then no longer suffers the three problems. Other researchers previously identified, qualitatively, the three problems focused upon in this article. What is new in this article is (i) the illustration of those problems with quantitative fits to robust, new data from simple (almost minimalist) human learning experiments, and (ii) an emphasis that the three problems exist simultaneously in standard backprop, so that solving one does not by itself make standard backprop a viable model of human category learning. Thus, the main goal of this article is to provide new emphasis and illustrations of these issues with quantitative fits to data from straightforward category learning experiments.

This article is organized as follows. I first describe the two models and point

out some general behavioral properties that can be gleaned from their structures. In the subsequent three sections, the models are applied to three situations in human category learning, demonstrating the three failures of standard backprop already mentioned. A modified version of backprop is then presented, which approximates ALCOVE and avoids the problems of standard backprop. The final section mentions other problems confronted by these models.

2. The Models

Both ALCOVE and standard backpropagation are feed-forward connectionist networks that learn by gradient descent on error. Thus, they both consist of a set of input nodes that encode the stimulus to be categorized, a set of output nodes that encode the category label, and a set of intermediate, 'hidden', nodes that transform the input representation into some internal representation. The layers of nodes are connected by weighted links, through which activation spreads from the input nodes, to the hidden nodes, to the category nodes. The models differ in two ways: (1) their particular choice of internal representation; and (2) whether or not there is a mechanism for dimensional attention learning.

2.1. ALCOVE

The architecture of ALCOVE was motivated by a molar-level psychological theory, Nosofsky's (1986) generalized context model (GCM). Like the GCM, ALCOVE assumes that input patterns can be represented as points in a multi-dimensional psychological space, as determined by multi-dimensional scaling (MDS) algorithms (Kruskal, 1964; Shepard, 1962). Thus, the first step in applying the model is determining the psychological coordinates of the stimuli. To do this, one obtains similarity ratings (or confusabilities) of pairs of stimuli, and determines the coordinate values in psychological space that best predict those similarities. This process is analogous to generating a spatial map of cities when all you are told is the distances (dissimilarities) between cities.

Each input node of ALCOVE encodes a single psychological dimension, with the activation of the node indicating the value of the stimulus on that dimension. Thus, if ψ_i is the psychological scale value of the stimulus on dimension i, then the activation of the ith input node is

$$a_i^{in} = \psi_i \tag{1}$$

Figure 1 shows the architecture of ALCOVE, illustrating a case with just two input dimensions.

Using MDS coordinates to encode the input does not introduce clandestine degrees of freedom into the model. Rather, the MDS solution is determined completely independently from the similarity ratings. Thus, the MDS representation *constrains* the model, and unlike some applications of standard backprop, we are not allowed to assume just any input representation that happens to be convenient. Of course, some other input representation might in fact work better, but the point remains that the MDS representation does not provide any extra degrees of freedom.[1]

The ith input node is gated by a dimensional *attention strength* α_i. The attention strength on a dimension changes to reflect the relevance of that dimension for the

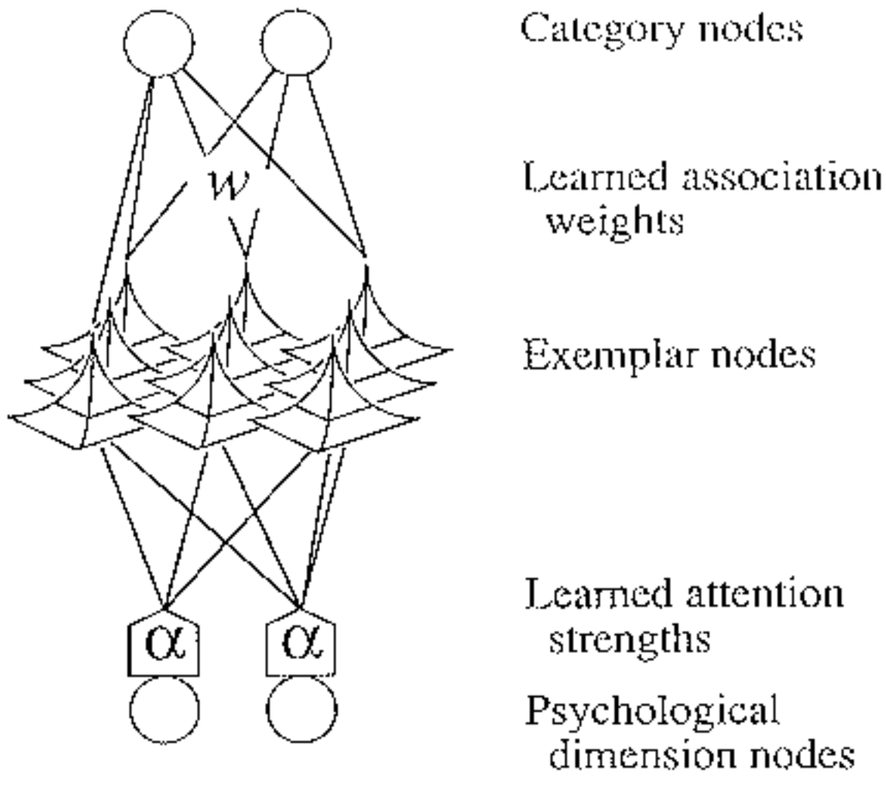


Figure 1. The structure of ALCOVE. The pyramids in the hidden layer indicate the activation profile of hidden nodes, as determined by equation (2), with $r = q = 1$.

particular categorization task at hand, as determined by a learning mechanism described below.

Each hidden node corresponds to a position in the multi-dimensional stimulus space, with one hidden node placed at the position of every training exemplar.[2] Each hidden node is activated according to the psychological similarity of the stimulus to the exemplar represented by the hidden node. The similarity function comes from the GCM and the work of Shepard (1962, 1987): Let the position of the jth hidden node be denoted $(h_{j1}, h_{j2}, \ldots)$, and let the activation of the jth hidden node be denoted a_j^{hid}. Then

$$a_j^{\text{hid}} = \exp\left(-c\left(\sum_i \alpha_i | h_{ji} - a_i^{\text{in}} |^r \right)^{q/r} \right) \tag{2}$$

where c is a positive constant called the *specificity* of the node, where the sum is taken over all input dimensions, and where r and q are constants determining the similarity metric and similarity gradient, respectively. For separable psychological dimensions, the city-block metric ($r = 1$) is used, while integral dimensions might call for a Euclidean metric ($r = 2$). An exponential similarity gradient ($q = 1$) is used here (Shepard, 1987).

When $r = 2$ in equation (2), the hidden nodes are a type of *radial basis function* (RBF), and ALCOVE can be construed as a type of radial basis function interpolation network (Broomhead & Lowe, 1988; Moody & Darken, 1989; Poggio & Girosi, 1990; Robinson *et al.*, 1988). Indeed, ALCOVE was born with the simple observation that the GCM can be implemented as an RBF network (Kruschke, 1990). Interestingly, a very similar model was independently invented by Hurwitz (1990), who was (quite differently) motivated by Estes's (1988) suggestion to combine exemplar representations with error-driven learning.

The dimensional attention strengths adjust themselves so that exemplars from different categories become less similar, and exemplars within categories become more similar. Consider a simple case of eight stimuli that form the corners of an octagon in a two-dimensional stimulus space, as shown in Figure 2. The stimuli are assigned to one of two categories, indicated by filled or open circles. Figure 2(a) shows a case in which one dimension can be ignored without loss of

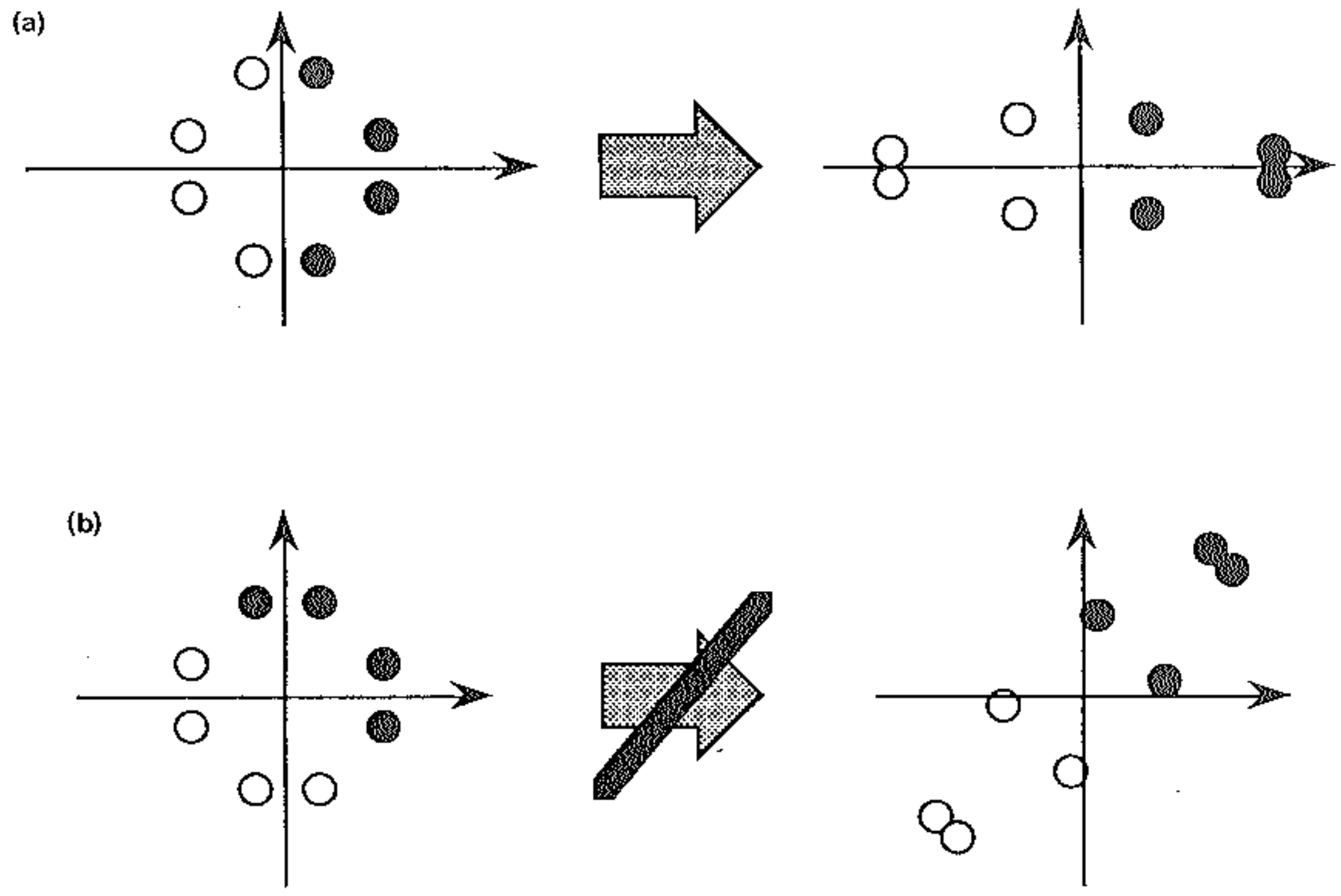


Figure 2. (a) Increasing attention to the horizontal dimension and decreasing attention to the vertical dimension causes exemplars of the two categories (denoted by filled and open circles) to have greater between-category dissimilarity and greater within-category similarity. (b) ALCOVE cannot differentially attend to diagonal axes.

classification accuracy. In this case, ALCOVE learns to increase the attention strength on the relevant dimension, and to decrease the attention strength on the irrelevant dimension. Increasing attention has the effect of stretching the dimension, and decreasing attention shrinks the dimension. Figure 2(b) shows a case in which neither dimension can be ignored without loss of classification accuracy. ALCOVE cannot stretch or shrink the stimulus space diagonally. As we will see, standard backprop does not share this anisotropy, and can differentially emphasize any direction in stimulus space. This difference has dramatic consequences for the models' predictions of the relative ease of learning such category structures, as will be demonstrated later.

Each hidden node in ALCOVE is connected to output nodes that correspond to response categories. The connection from the jth hidden node to the kth category node has a connection weight denoted w_{kj}, called the *association weight* between the exemplar and the category. The category nodes are activated by the linear rule used in the GCM and in the network models of Gluck and Bower (1988):

$$a_k^{\text{out}} = \sum_{\substack{\text{hid} \\ j}} w_{kj} a_j^{\text{hid}} \tag{3}$$

Category activations are mapped to response probabilities using the same choice rule (Luce, 1963) as was used in the GCM and network models:

$$\Pr(K) = \exp(\varphi a_K^{\text{out}}) / \sum_{\substack{\text{out} \\ k}} \exp(\varphi a_k^{\text{out}}) \tag{4}$$

where φ is a scaling constant. In other words, the probability of classifying the given stimulus into category K is determined by the magnitude of category K's activation relative to the sum of all category activations.

Suppose the model is applied to the situation illustrated in Figure 2. In this case, there are two psychological dimensions, hence two input nodes; eight training exemplars, hence eight hidden nodes; and two categories, hence two output nodes. When an exemplar is presented to ALCOVE, the input nodes are activated according to the component dimensional values of the stimulus (equation (1)). Each hidden node is then activated according to the similarity of the stimulus to the exemplar represented by the hidden node, using the attentionally weighted metric of equation (2). Thus, hidden nodes near the input stimulus are strongly activated, and those farther away in psychological space are less strongly activated. Then the output (category) nodes are activated by summing across all the hidden (exemplar) nodes, weighted by the association weights between the exemplars and categories, as in equation (3). Finally, response probabilities are computed using equation (4).

The dimensional attention strengths, α_i, and the association weights, w_{kj}, are learned by gradient descent on sum-squared error, as used in standard backprop (Rumhelhart *et al.*, 1986a) and in the network models of Gluck and Bower (1988). Each presentation of a training exemplar is followed by feedback indicating the correct response. The feedback is coded in ALCOVE as *teacher* values, t_k, given to each category node. For a given training exemplar and feedback, the *error* generated by the model is defined as

$$E = \frac{1}{2} \sum_{\substack{\text{out} \\ k}} (t_k - a_k^{\text{out}})^2 \tag{5}$$

where the teacher values are defined as

$$t_k = \begin{cases} \max(+1, a_k^{\text{out}}) & \text{if stimulus} \in k \\ \min(-1, a_k^{\text{out}}) & \text{if stimulus} \notin k \end{cases} \tag{6}$$

These teacher values are defined so that activations 'better than necessary' are not counted as errors. Thus, if a given stimulus should be classified as a member of the kth category, then the kth output node should have an activation of *at least* +1. If the activation is greater than +1, then the difference between the actual activation and +1 is not counted as an error. Because these teacher values do not mind being outshone by their students, I call them *humble* teachers. The motivation for using humble teacher values is that the feedback given to subjects is nominal, indicating only which category the stimulus belongs to, and not the degree of membership. Hence the teacher used in the model should only require some minimal level of category-node activation, and should not require all exemplars to produce ultimately the same activations. Humble teachers are discussed further by Kruschke (1990, 1992), and they do not play a central role in this article.

Upon presentation of a training exemplar to ALCOVE, the association weights and attention strengths are changed by a small amount so that the error decreases. Following Rumelhart *et al.*, they are adjusted proportionally to the (negative of the) error gradient, which leads to the following learning rules, for $r = q = 1$ (derived in Kruschke, 1990, 1992):

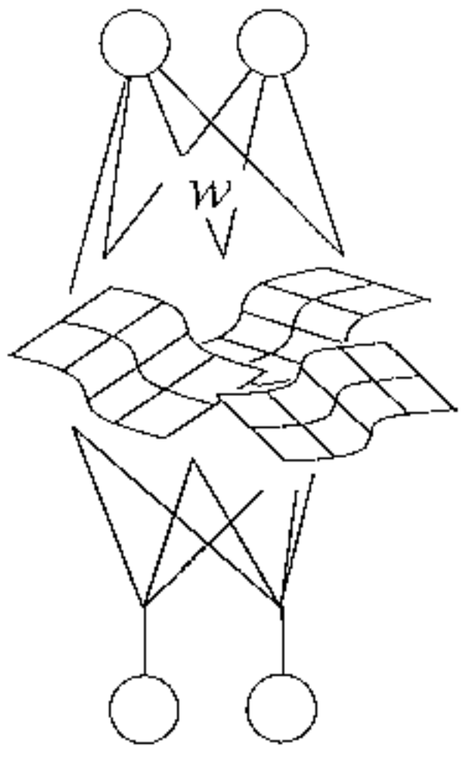


Figure 3. The structure of standard backpropagation.

$$\Delta w_{kj}^{\text{out}} = \lambda_w (t_k - a_k^{\text{out}}) a_j^{\text{hid}} \tag{7}$$

$$\Delta \alpha_i = -\lambda_\alpha \sum_{\substack{\text{hid} \\ j}} \left(\sum_{\substack{\text{out} \\ k}} (t_k - a_k^{\text{out}}) w_{kj} \right) a_j^{\text{hid}} c \left| h_{ji} - a_i^{\text{in}} \right| \tag{8}$$

where the λs are constants of proportionality ($\lambda > 0$) called 'learning rates'. The same learning rate, λ_w, applies to all the output weights. Likewise, there is only one learning rate, λ_α, for all the attention strengths. If application of equation (8) gives an attention strength a negative value, then that strength is set to zero, because negative values have no psychologically meaningful interpretation.

Learning in ALCOVE proceeds as follows: For each presentation of a training exemplar, activation propagates to the category nodes as described previously. Then the teacher values are presented and compared with the actual category-node activations. The association weights and attention strengths are then adjusted according to equations (7) and (8).

In fitting ALCOVE to human learning data, there are four free parameters: the fixed specificity c in equation (2); the probability mapping constant ϕ in equation (4); the association weight learning rate λ_w in equation (7); and the attention strength learning rate λ_α in equation (8).

2.2. *Standard Backpropagation*

Standard backprop (Figure 3) uses *linear-sigmoid* nodes in its hidden layer, which have activation determined by

$$a_j^{\text{hid}} = 1 / \left(1 + \exp \left[-g \left(\sum_{\substack{\text{in} \\ i}} w_{ji}^{\text{hid}} a_i^{\text{in}} - \theta_j \right) \right] \right) \tag{9}$$

where g is a constant called the *gain* of the node (Kruschke & Movellan, 1991) and θ_j is the *threshold* of the node. The linear-*sigmoid* function was motivated as a generalized, or smoothed, version of the linear-*threshold* function in 'neuron-like' perceptrons.

The activation profiles of hidden nodes in ALCOVE and in backprop, as determined by equations (2) and (9), are shown in Figure 4. Note that the level

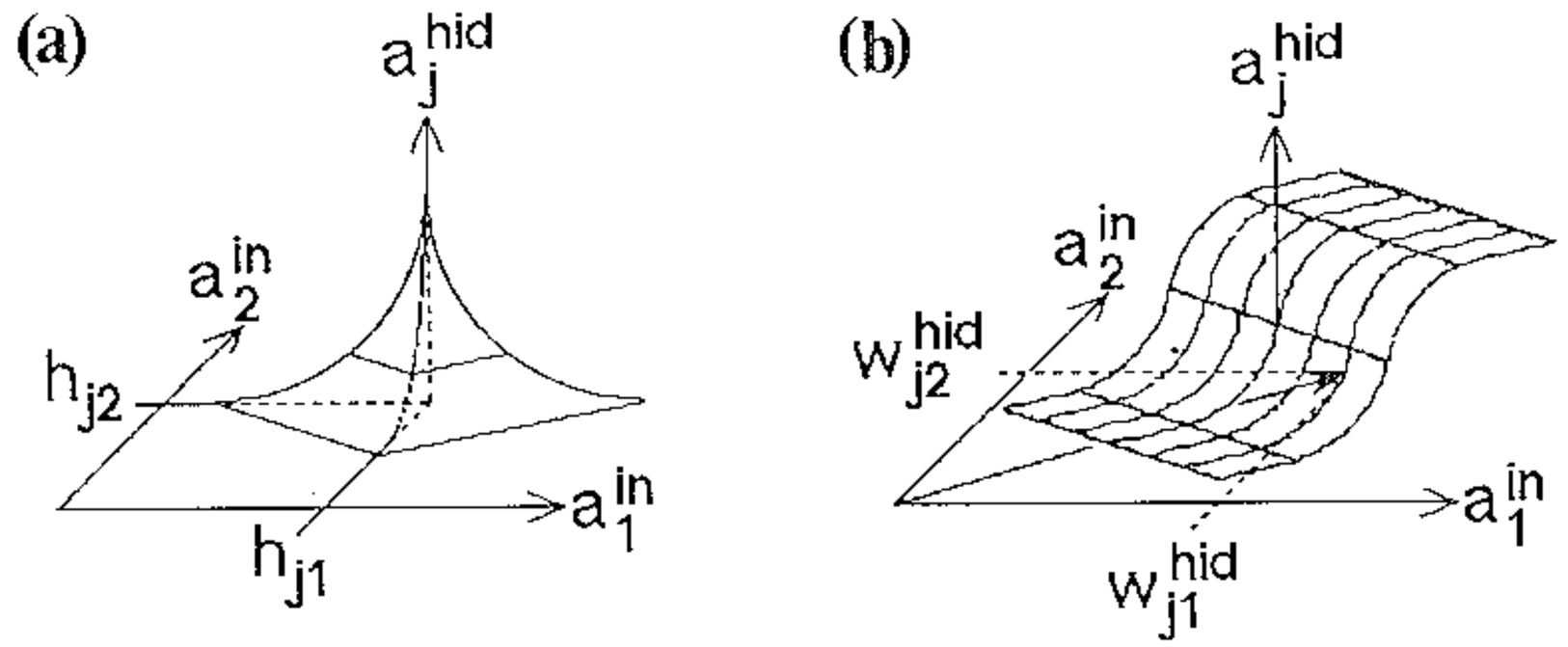


Figure 4. (a) Activation profile of a hidden node in ALCOVE (equation (2), with $r = q = 1$). (b) Activation profile of a hidden node in standard backpropagation (equation (9)). Hidden nodes in ALCOVE have a localized receptive field, whereas hidden nodes in backprop respond to an entire half-space.

contours of the ALCOVE node are iso-distance contours (diamond shaped for a city-block metric), whereas the level contours of the backprop node are linear (hyperplanes). Examples of level contours are shown in Figure 4 by the lines that mark 'horizontal' cross-sections through the activation profiles. In backprop, the *weights* in equation (9) determine the *orientation* of the linear level-contour in input space; the *threshold* in equation (9) determines the *distance* of the $a_j^{hid} = 0.50$ level contour from the origin; and the *gain* in equation (9) determines the *steepness* of the sigmoidal hill.

One important difference between the structures of ALCOVE and backprop is that the linear level-contours of the backprop node can be oriented in any direction in input space (depending on the hidden weights), whereas attention learning in ALCOVE can only stretch or shrink along the given input dimensions (recall the discussion accompanying Figure 2). In backprop, the linear level-contours can be equally easily aligned along the vertical or diagonal category boundaries in Figure 2. Consequently, backprop shows virtually no difference in learning speed between the two categorizations, as will be described at greater length below.

The gain parameter is not usually included in backpropagation (i.e. it is usually fixed at $g = 1.0$), but is included here for two reasons. First, it plays a role comparable to the specificity parameter in ALCOVE. Just as specificity determines the steepness of the generalization gradient for hidden nodes in ALCOVE, gain determines the steepness of the sigmoidal hill for hidden nodes in backdrop. Second, it determines the effective range for initializing the random weights and thresholds of the hidden nodes. That is, the initial values of the weights and thresholds in equation (9) are drawn randomly from a uniform distribution on the interval $[-1, +1]$, and the gain acts as a multiplier to make the effective range $[-g, +g]$ (as can be seen by distributing g over the w_{ji}^{hid} and θ_j terms in equation (9). This is important because it has been shown that the magnitude of initial weights can affect the behavior of backprop (Kolen & Pollack, 1990).

In order to make backprop and ALCOVE comparable in their output assumptions, the backprop network is given linear output nodes (equation (3)) with response probabilities determined by equation (4). The output weights are initialized to zero, for the same reason as in ALCOVE, viz, that initially there should be

no particular correspondence of stimuli (or their re-representation in the hidden layer) with category labels. The input layer of backprop is also assumed to use the same representation used by ALCOVE, because backprop makes no particular assumptions about the input presentation. The upshot is that there are two critical distinctions between ALCOVE and backprop: the difference in hidden node activation functions; and the presence or absence of a dimensional attention learning mechanism.

Learning in backprop proceeds as follows. Upon presentation of a training exemplar, the weights and thresholds are adjusted proportionally to the (negative of the) error gradient, which leads to the following learning rules (for derivations see Kruschke & Movellan, 1991; Rumelhart *et al.*, 1986a):

$$\Delta w_{kj}^{\text{out}} = \lambda_{\text{out}}(t_k - a_k^{\text{out}})a_j^{\text{hid}} \tag{10}$$

$$\Delta w_{ji}^{\text{hid}} = \lambda_{\text{hid}}\left(\sum_{\substack{\text{out}\\k}}(t_k - a_k^{\text{out}})w_{kj}^{\text{out}}\right)(1 - a_j^{\text{hid}})a_j^{\text{hid}}ga_i^{\text{in}} \tag{11}$$

$$\Delta\theta_j = -\lambda_\theta\left(\sum_{\substack{\text{out}\\k}}(t_k - a_k^{\text{out}})w_{kj}^{\text{out}}\right)(1 - a_j^{\text{hid}})a_j^{\text{hid}}g \tag{12}$$

The same learning rate, λ_{out}, applies to all the output weights. Likewise, there is only one learning rate, λ_{hid}, for all the hidden weights, and one learning rate, λ_θ, for all the hidden thresholds. Note that equation (10) is the same as equation (7), because the output layers of the two models are assumed to have the same structure. Note also the similarity of equations (11) and (12) to equation (8).

In fitting backprop to data there are five parameters: the three learning rates in equations (10), (11) and (12); the gain g, and the choice probability constant ϕ.

2.3. Summary of Models

Throughout this article, when I use the term 'backprop' I am referring to the use of linear-sigmoid hidden nodes as in standard backpropagation. When I mean to refer to the learning mechanism, I will call it gradient descent on error, or error-driven learning. The target of this article is linear-sigmoid hidden nodes and dimensional attention, not gradient descent on error.

ALCOVE can be construed as a coalescence of three intellectual currents. First, it uses an exemplar-based internal representation that stems directly from theories of similarity-based generalization and attentionally weighted stimulus dimensions proposed by Shepard (1957, 1987), Medin and Schaffer (1978), Estes (1986) and Nosofsky (1986). Second, it uses error-driven learning, as in the models of Rescorla and Wagner (1972), Rumelhart *et al.* (1986a/b), and Gluck and Bower (1988). Third, it is essentially a form of radial-basis function interpolation network, as described by Broomhead and Lowe (1988), Robinson *et al.* (1988), Moody and Darken (1989), Poggio and Girosi (1990), and others. Thus ALCOVE should not be construed as a model opposed to standard backprop, but rather as a variant of it.

3. Lack of Selective Attention

As described in the previous section, backprop and ALCOVE differ in two critical ways: the shapes of their hidden node receptive fields, and the presence or absence of learned dimensional attention strengths. This section will concentrate on the dimensional attention strengths, and subsequent sections will focus on the receptive fields.

3.1. Filtration vs Condensation

Imagine attending a display of fireworks, and hearing the crowd respond "ooh!" to some and "ahh!" to others. Let's suppose that the responses are not random, but are determined by some visible properties of each burst, and that it is now our task to *learn* which bursts get which response. Suppose that all red fireworks elicit an "ooh!", while all other colors conjure an "ahh!" It is intuitively plausible that one would quickly learn to attend to the dimension of color, and ignore other stimulus dimensions. Suppose instead that the cheer for a burst could only be determined by the combination of two (or more) dimensions, such as color and size. It seems likely that accurate classification learning would take longer. The first situation, in which categorization of a stimulus could be accomplished by attending to just a subset of the available stimulus dimensions, has been called a 'gating' or 'filtration' task, because the irrelevant dimensions can be filtered or gated away, without loss of classification accuracy. The second situation, in which more than one dimension must be attended for accurate classification, has been called a 'condensation' task, because information from more than one dimension must be condensed into a single classification decision (Garner, 1974; Posner, 1964).

It has been well established that filtration tasks are indeed easier than condensation tasks (e.g. Garner, 1974; Gottwald & Garner, 1972, 1975; Kemler & Smith, 1978; Posner, 1964), confirming the intuition in the hypothetical case of classifying fireworks. The advantage of filtration over condensation can be appraised as a robust and fundamental phenomenon that models of category learning should address.

Explanations of filtration advantage have invoked the notion of selective attention. For example, Garner (1974) argued that condensation was difficult because the subject could not attend to a combination of dimensions with the same efficiency as he or she could (selectively) attend to a single dimension. It seems only natural, then, that models of category learning should include mechanisms for selective attention to stimulus dimensions. In this section I show that backprop does not have an appropriate form of selective attention, and fails to show filtration advantage. On the other hand, ALCOVE has dimensional attention strengths built in, and does exhibit filtration advantage.

In order to compare directly the models' quantitative predictions, new data had to be obtained from filtration and condensation categorization tasks. Unfortunately, pyrotechnic displays are unwieldy stimuli, so instead I used category exemplars consisting of rectangles that varied in height, with an interior segment that varied in its lateral position (Figure 5). Different groups of subjects were trained on the four category structures shown in Figure 6. Only eight different stimuli were presented, four of which were assigned to one category and the remaining four to the other category, indicated in Figure 6 by blank and filled circles. There were two filtration conditions, one in which height was relevant and

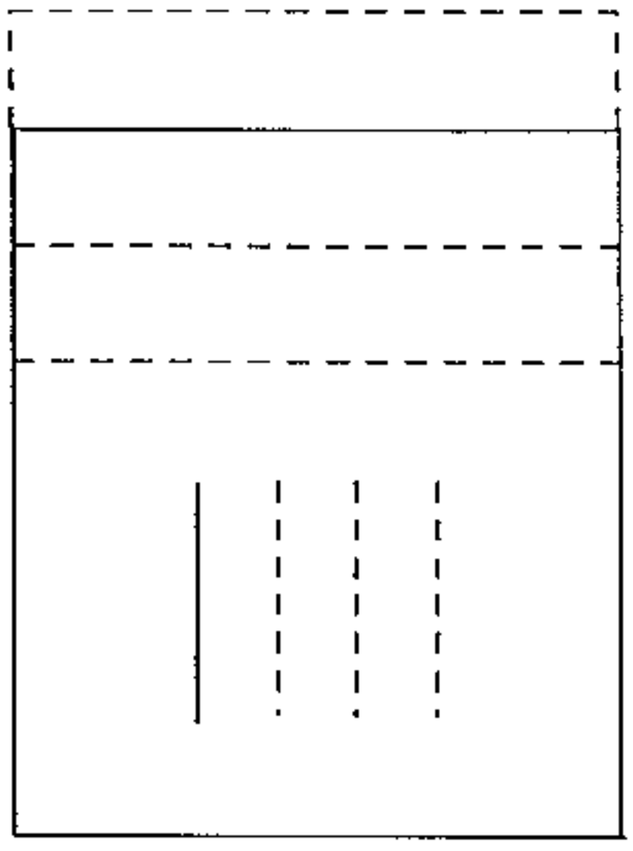


Figure 5. Stimuli. Solid lines show one combination of rectangle height and lateral position of interior segment. Dotted lines show alternative heights and positions.

one in which position was relevant (Figure 6(a)), and two condensation conditions (Figure 6(b)).

I chose the structures in Figure 6 primarily because they are (very nearly) rotations of each other, and standard backprop is insensitive to rotations of the input space. A second motive for these structures was that the clustering of exemplars, considered alone, predicts that the condensation situations should be at least as easy the filtration conditions. Exact quantitative predictions on the basis of clustering depend on how one wishes to define a measure of clustering. As an illustration, suppose we use the mean city-block separation of exemplars within categories vs between categories (Medin & Schwanenflugel, 1981), and let the

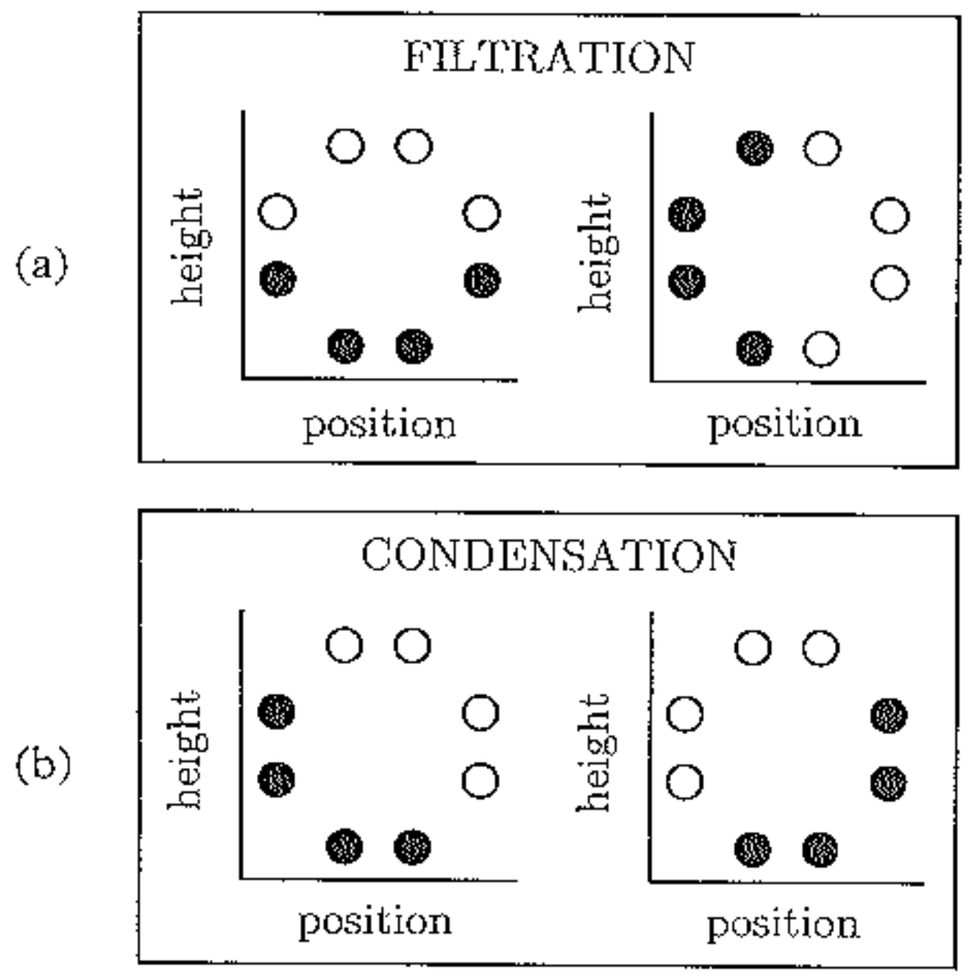


Figure 6. Structure of the filtration and condensation categories. Open circles denote one category, filled circles the other.

distance between levels on each dimension be 1 scale unit. Then both the filtration and condensation structures have a mean within-category separation of 2.33 and a mean between-category separation of 3.25. If city-block separations are (non-linearly) converted to similarities using the exponential function of distance in equation (2) (with $c = r = q = 1$), then the condensation structure should be easier than the filtration structure, as the mean similarity of exemplars within categories is 0.16 for condensation but only 0.13 for filtration, and the mean similarity of exemplars between categories is only 0.049 for condensation but 0.074 for filtration. A third consideration in the choice of these structures was that the same exemplars are used for both the filtration and condensation tasks; only the category assignment changes. Thus any differences between the filtration and condensation categories can be attributed to their structures rather than to effects of individual exemplars.

3.1.1. Procedure. Stimuli were presented with a PC-clone computer using VGA resolution, as white lines against a black background. Viewing distance was about 0.9 m, so that the height of the tallest rectangle subtended about 13 degrees of visual angle. The rectangles were presented so that the lower horizontal line was in the same position on every trial, centred horizontally on the screen. Of the 16 possible combinations of dimension values, only eight were used, corresponding to the abstract structure in Figure 2. In all experiments reported in this article, subjects were run individually in a dimly lit, quiet booth. All experiments were programmed using Micro Experimental Laboratory (Schneider, 1988).

Instructions were presented to the subject on the computer screen, and read aloud by the experimenter. Subjects were told that they must learn which stimulus belonged to which category. The category labels were 'B' and 'N'. Subjects responded by using the index and middle fingers of their dominant hand to press the corresponding keys on the keyboard. The instructions said that the stimuli varied on just two dimensions, height and position. As part of the instructions, subjects were shown all the stimuli without any category feedback (and without any responses from the subject). Each of the eight stimuli was shown twice, in a random sequence that was the same for all subjects. Subjects were instructed that there was no emphasis on response speed, and that they had up to half minute to respond on each trial.

Each training trial consisted of a presentation of a stimulus, which was terminated when the subject pressed a response key, followed by 1000 ms feedback indicating whether the response was 'correct' or 'wrong' accompanied by a 333 ms tone if wrong, followed by 750 ms feedback indicating the correct response ('That was a B' or 'That was an N').

The four category distinctions were given to different groups of subjects. Category labels were counterbalanced within groups, so that a given stimulus had correct label 'B' for half the subjects and 'N' for the other half. Every subject in every group saw the same fixed sequence of stimuli; all that varied between groups was the category labels assigned to the stimuli. There were 64 uninterrupted training trials. The experiment lasted about 20 minutes.

3.1.2. Subjects. A total of 160 subjects participated, 40 in each category type, for partial credit in an introductory psychology course.

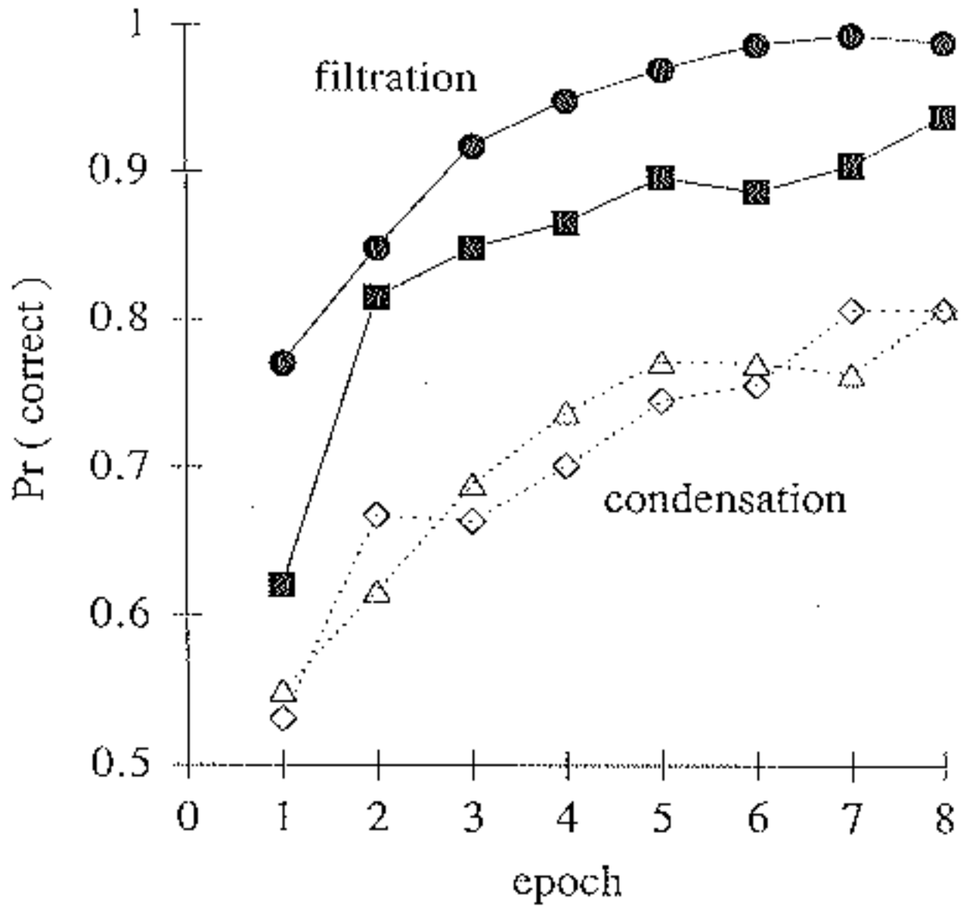


Figure 7. Human learning data for the four category structures shown in Figure 6. One 'epoch' is one sweep through the eight different stimuli. Filled circles show the position-relevant filtration; filled squares show the height-relevant filtration. Open markers show results from the two condensation conditions.

3.1.3. Results. Results are summarized in Figure 7. Each datum shows the mean per cent (Pr) correct for the preceding eight trials (one 'epoch'). Two effects are evident from visual inspection alone: filtration categorizations (filled markers and solid lines) are learned much faster than condensation categorizations (open markers and dotted lines); and, of the two filtration categorizations, the one with position relevant (filled circles) was learned faster than the one with height relevant (filled squares).

Inferential statistics confirm the reliability of those differences. To perform the analyses, the 64 trials were divided into four bins of 16 trials each, and the proportion of trials correct was computed for each subject in each bin. To better approximate homogeneous variances, the proportions were transformed using $y = 2 \arcsin \sqrt{x}$ (Winer *et al.*, 1991, p. 356). Thus, trial-bin was a within-subjects factor with four levels and category type was a between-subjects factor with four levels. Bonferroni T-tests at each level of trial-bin confirmed that position-relevant filtration was learned faster than height-relevant filtration, which was learned faster than either condensation, which were not significantly different. MANOVA showed no significant interaction of trial-bin by category type, $F(9, 374.94) = 1.5875$, $P = 0.1170$ using Wilks's lambda, and $F(9, 468) = 1.5757$, $P = 0.1197$ using the Pillai–Bartlett trace.

The main point is that acquisition of the filtration categorizations is *faster* than for condensation, not that ultimate performance on the condensation task is worse, for there is little doubt that subjects could have eventually achieved asymptotic performance in the condensation task nearly equal to that in the filtration condition if they were trained long enough. These results agree with previous results in the literature (Garner, 1974; Gottwald & Garner, 1972; Posner, 1964).

Again, the reason for running these new experiments was not to demonstrate new qualitative phenomena, but to obtain new trial-by-trial learning curves to

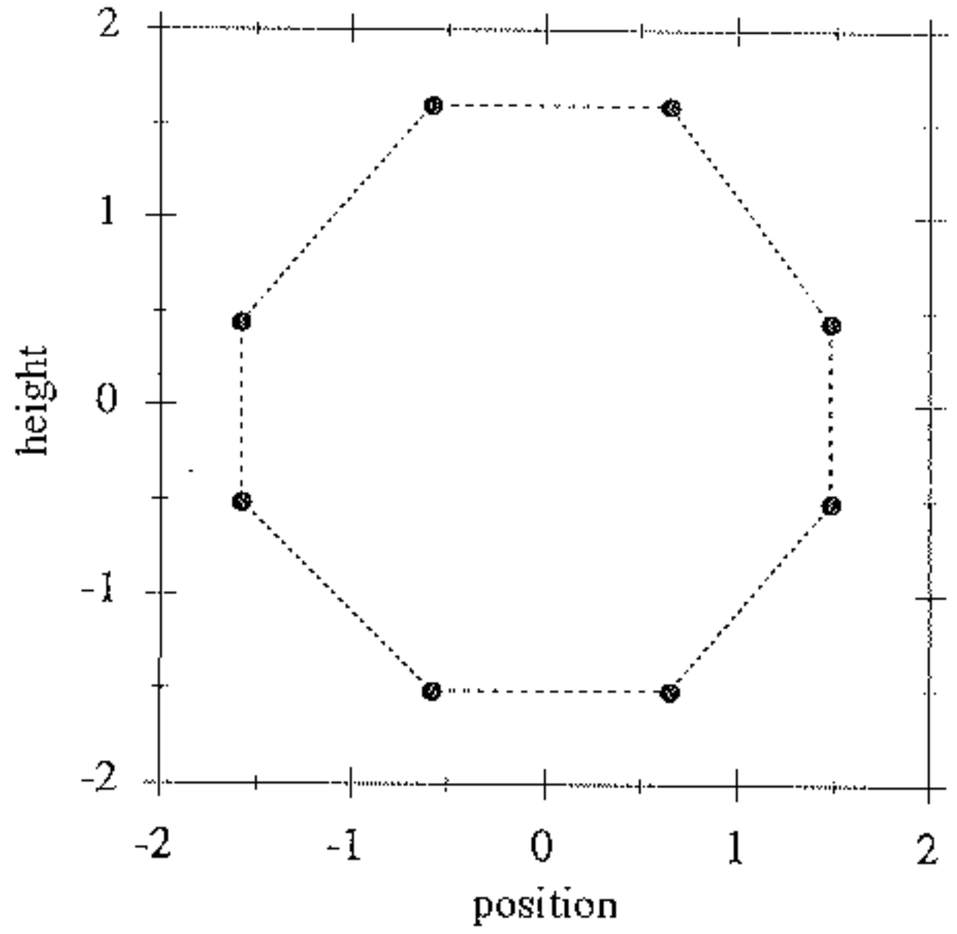


Figure 8. Locations of stimuli in psychological space. Stimulus points are connected by dotted lines to emphasize the departure of the psychological locations from the equal-increment physical distribution.

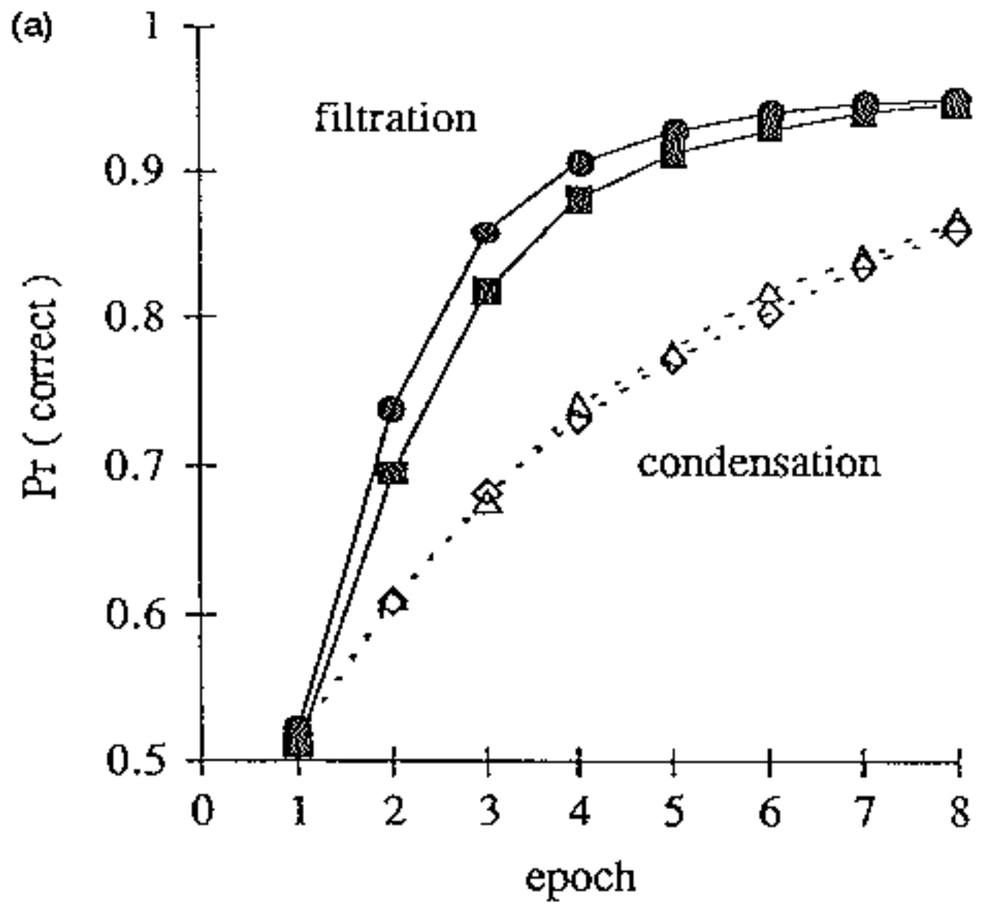
which the models could be fitted. The particular design employed here has the advantages of using the same stimuli for both filtration and condensation conditions, and of making the condensation task at least as easy as filtration when considering clustering alone.

3.2. Fit of ALCOVE

ALCOVE requires the stimuli to be encoded by their psychological coordinates, so a similarity-scaling study was run. Details of the procedure and scaling solution are presented in the appendix. The best-fitting psychological coordinates of the stimuli are shown in Figure 8. The two dimensions were about equally salient, in that the total range is about the same on the two dimensions. Thus, salience alone should not play a major role in determining the ease of categorizing by one dimension or the other. The *physical* values of height and position were equally spaced (see Figure 5), but the *psychological* values were not. In particular, the middle interval on the position dimension was psychologically bigger than the middle interval of the height dimension. That implies that a category boundary dividing left and right positions should be easier to learn than a category dividing upper and lower heights, just as was observed in the learning data.

Figure 9 shows the best fit of ALCOVE to the human learning data. ALCOVE had two input nodes (one for each psychological stimulus dimension), eight hidden nodes (one per training exemplar), and two output nodes (one for each category). The stimuli were encoded on the input nodes by their psychological scale values. All four learning curves were fitted simultaneously, using the same four parameter values. The best fit to the trial-by-trial (not epoch-by-epoch) learning data produced a root mean squared deviation (RMSD) of 0.116, using parameter values of $\phi = 1.568$, $c = 1.662$, $\lambda_w = 0.08431$ and $\lambda_\alpha = 0.6593$.

ALCOVE clearly shows the two main phenomena seen in the human data:



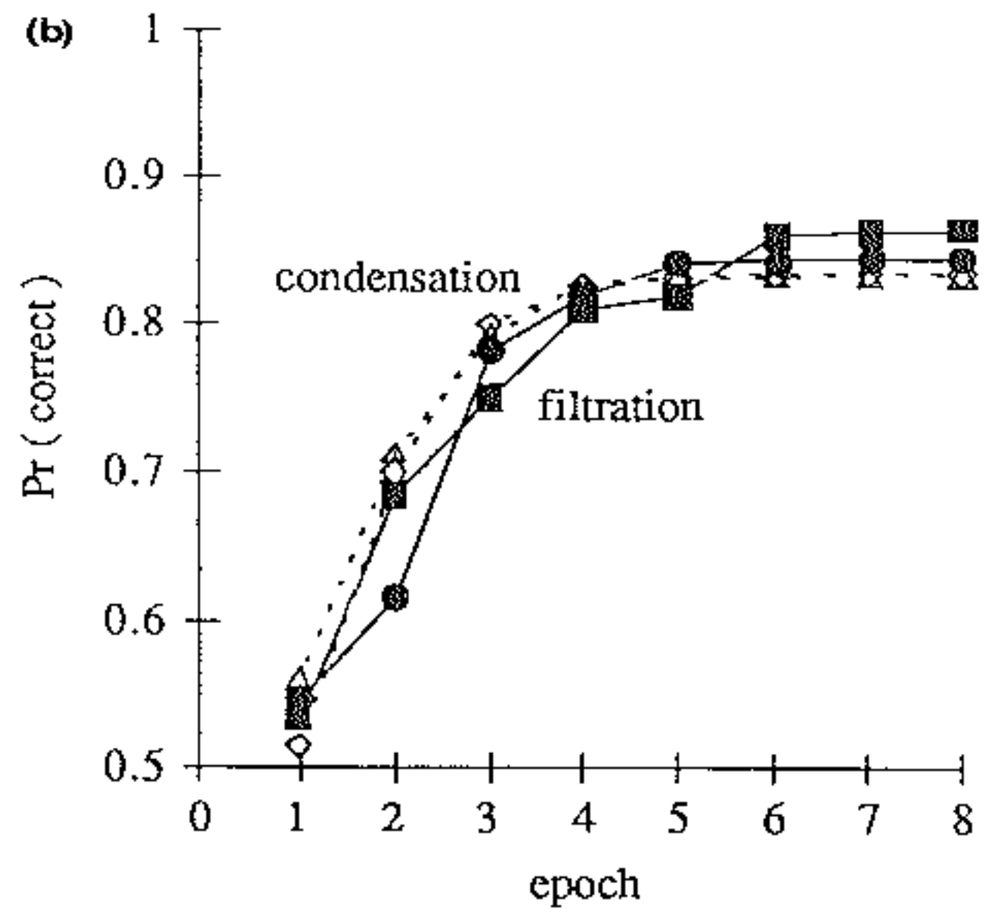


Figure 9. Best fit of (a) ALCOVE and (b) backprop to data shown in Figure 7.

filtration is learned much faster than condensation, and the position dimension is learned faster than the height dimension.[3] The reason it is capable of showing filtration advantage was described in conjunction with Figure 2: the dimensional attention strengths can facilitate performance on the filtration tasks, but cannot yield much benefit in the condensation tasks. In the condensation condition, ALCOVE still makes adjustments to the dimensional attention strengths, but those adjustments cannot *differentially* affect the two diagonal axes.

To understand why ALCOVE learns the height-relevant filtration more slowly than the position-relevant dimension, recall that exemplar nodes are activated according to their proximity to the current stimulus. Thus, any single stimulus will fully activate its corresponding exemplar node, and partially activate all the other examplar nodes (some only negligibly). As can be gleaned from the learning

equation for association weights (equation (7)), on any given trial *all* the association weights to a given category node will be changed in the *same* direction (positive or negative), regardless of the true category assignments of the various exemplars. The magnitude of change is proportional to the proximity of the exemplar to the current stimulus, as reflected in the activation of the node. That action benefits other exemplars in the same category, but harms exemplars in other categories. The farther away the other-category exemplars are from the current stimulus, the less is the harm to them. Because the intermediate interval of the position dimension is larger than the intermediate interval of the height dimension, the position-relevant filtration suffers less harm, and is therefore faster, than the height-relevant filtration.

3.3. Fit of Backprop

Backprop was fitted to the data using a network architecture matched as closely as possible to that of ALCOVE. Eight hidden nodes were used, the same number as were used in ALCOVE. In principle, backprop could accurately learn any one of the four category distinctions using just two hidden nodes. (If the output nodes had threshold terms, or if the hidden nodes had activation values in the interval $[-1, +1]$ instead of $[0, +1]$, then only one hidden node would be needed.)

For any choice of parameter values, the fit was measured in the same way as for ALCOVE, but choice predictions of backprop were computed by first averaging over 500 different random initializations of the hidden weights and thresholds. The same 500 initializations were used for each category type.

The best fit of backprop to the learning data is shown in Figure 9. The best fit yielded an RMSD of 0.152, by using $\phi = 0.6636$, $\lambda_{out} = 0.2049$, $\lambda_{hid} = 1.091$, $\lambda_\theta = 0.04159$ and $g = 2.249$. The qualitative behavior of backprop departs badly from the data. Indeed, backprop learns filtration and condensation at essentially the same pace. The best backprop can do is to try to match the mean learning curve across all four category types (except for the initial trials, where it, like ALCOVE, cannot learn fast enough).

Why did backprop fail to perform differently on the filtration and condensation tasks? As explained earlier, in conjunction with Figure 4, the weights leading into a hidden node determine the orientation of its linear level-contour. Those level contours can align with the diagonal category boundaries in the condensation tasks just as easily as they can align with the canonical axes in the filtration task. Standard backpropagation learning is isotropic in that sense, unlike human learning. In more anthropomorphic terms, backprop is not constrained to attend to the psychological dimensions as given in the input representation, but can just as easily attend to new dimensions defined as linear combinations of the input dimensions. Apparently people cannot do that so easily.

Backprop does not lack selective attention utterly, but lacks proper constraints on selective attention. Backprop can selectively attend to any direction in stimulus space, whereas ALCOVE can selectively attend only to the given psychological dimensions.

3.4. Extended Backprop

I have suggested that the reason backprop fails to learn condensation more slowly than filtration is that its hidden nodes can orient themselves in any direction in

input space. Perhaps the problem can be solved quite simply by importing the lower layer of ALCOVE into backprop. That is, each input node should be gated by an attention parameter α_i, and the weights leading into the hidden nodes should be fixed at their random initial values, so that they cannot re-orient themselves to accentuate the diagonal directions in stimulus space.[4]

In the extended version of backprop, the hidden node activation function is given by

$$a_j^{\text{hid}} = 1/\left(1 + \exp\left[-g\left(\sum_{\substack{\text{in}\\ i}} w_{ji}^{\text{hid}} \alpha_i a_i^{\text{in}} - \theta_j\right)\right]\right) \tag{13}$$

Compare equation (13) with equation (9); the only difference is the appearance of the attention strengths, α_i. In the extended model, we trade hidden weight learning for attention strength learning. So the learning rules in equations (11) and (12) no longer apply, but a new rule for attention strength learning is used:

$$\Delta\alpha_i = \lambda_\alpha \sum_{\text{hid}_j}\left(\sum_{\text{out}_k}(t_k - a_k^{\text{out}}) w_{kj}^{\text{out}}\right)(1 - a_j^{\text{hid}}) a_j^{\text{hid}} g w_{ji}^{\text{hid}} a_i^{\text{in}} \tag{14}$$

Compare equation (14) with the corresponding rule in ALCOVE, equation (8), and it can be seen that their basic structures are very similar. As in ALCOVE, the attention strengths are clipped at zero, because negative attention values have no clear meaning.

The extended backprop model has four free parameters: the response probability scaling constant ϕ, the output weight learning rate λ_{out}, the gain g and the attention strength learning rate λ_α. In fitting the model to data, the attention strengths are initialized as in ALCOVE, to values of $1/N$, where N is the number of stimulus dimensions.

To fit extended backprop to the data, the average of 500 different random initializations was taken as the predictions of the model, and the summed-squared deviation between the predicted and empirical choice probabilities was minimized. Figure 10 shows that the fit of the extended model is noticeably better than the basic model. The best fit had an RMSD of 0.125, with parameter values of $\phi = 1.46$, $\lambda_{\text{out}} = 0.0715$, $\lambda_\alpha = 4.80$ and $g = 13.5$. The filtration curves lie well above the condensation curves. (The slight advantage for the height-relevant filtration over the position-relevant filtration eludes explanation at this time.)

3.5. Discussion

The psychological phenomenon of filtration advantage was well established by previous researchers. The present study has two novelties. A novelty in experimentation was the use of simple category structures specifically designed to challenge standard backprop. A novelty in modeling was the demonstration that standard backprop cannot show filtration advantage, but an extended version, which includes dimensional attention learning, can.

The present experimental design, using category structures in which clustering favors condensation, also poses a strong challenge to some other models of categorization that do not incorporate dimensional attention learning. In particular, the rational model of Anderson (1991), the Cauchy-node model of Hanson and Gluck (1991), and the consequential-region model of Shanks and Gluck (1991), are extremely unlikely to show filtration advantage, in general. The

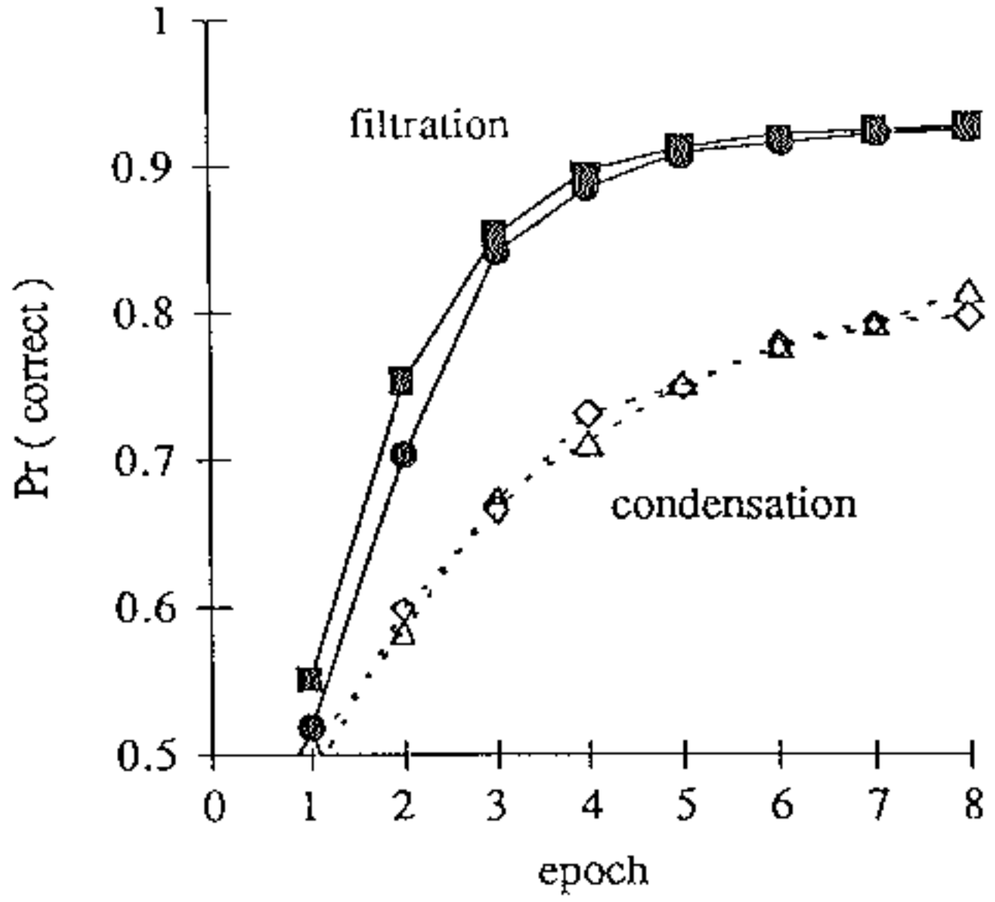


Figure 10. Best fit of extended backpropagation to human learning data of Figure 7. There is a strong filtration advantage.

stimulus structure used here is not necessarily optimal for testing all those models, however. The reason is that the filtration and condensation structures in Figure 6 differ not only in the orientation of their category boundaries, but also in their within-category dimensional variances. For example, consider the position-relevant filtration structure. The within-category variance on the position dimension is relatively small, whereas for the height dimension it is relatively large. On the other hand, for the condensation structures, the within-category dimensional variances are the same intermediate magnitude for both dimensions (except for small psychological scale differences). The rational model would be able to take advantage of that confound, and learn the filtration task more quickly because it has mechanisms to take advantage of the smaller within-category variances on the relevant dimension. Figure 11 shows alternative filtration and condensation structures which have the same within-category dimensional variances, and which, I suspect, would yield filtration advantage for humans, but not for the rational model.

The stimuli in these experiments varied on separable, not integral, dimensions. The claim that they were separable is supported primarily by the fact that a city-block distance metric fitted the similarity judgments better than a Euclidean metric, as reported in the appendix (Garner, 1974; Shepard, 1964). The results

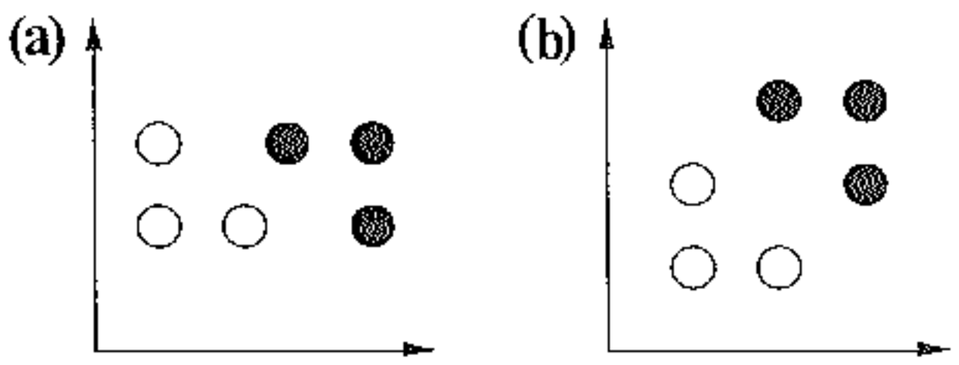


Figure 11. Stimulus configuration for testing the rational model: (a) filtration and (b) condensation.

and conclusions reported here are not necessarily restricted to separable-dimension stimuli, however. The advantage of filtration over condensation has also been observed for integral-dimension stimuli, although the difference is less robust. For example, some researchers have studied category learning using Munsell colors that varied on the integral dimensions of brightness (value) and saturation (chroma). Gottwald and Garner (1975), using a speeded sorting task of color cards, found that condensation was more difficult than filtration. Nosofsky (1987) found that a 'pink-brown' category distinction, a condensation task, was more difficult to learn than a 'saturation' or 'brightness' distinction, both filtration tasks. (Note, however, that Nosofsky's experiment was not explicitly motivated by considerations of filtration vs condensation.) Generalizing beyond the particular stimulus arrangements used in those experiments is somewhat risky: Gottwald and Garner (1975) used two condensation conditions, one of which had more stimuli than the filtration condition, and one of which was non-linear, so that the apparent filtration advantage might have been caused by one of those confounded factors. The condensation condition in Nosofsky's (1987) experiment also used more stimuli than the filtration conditions, and so the apparent filtration advantage might have been caused by differences in clustering or memory load. Nevertheless, it is not unreasonable to suppose a filtration advantage exists for integral-dimension stimuli, although further research using a stimulus structure such as that used here would be informative. ALCOVE is also able to show filtration advantage for integral dimensions, i.e. when a Euclidean distance metric is used ($r = 2$ in equation (2)). Of course, ALCOVE need not necessarily show filtration advantage, especially when the attentional learning rate is small.

4. Catastrophic Forgetting

Despite its ability to capture the difference between filtration and condensation, it is shown in this section that extended backprop inherits the problem of *catastrophic forgetting* from standard backprop (McCloskey & Cohen, 1989; Ratcliff, 1990).

Catastrophic forgetting in backprop is the phenomenon that previously learned associations can be largely forgotten when the network is trained on new associations (McCloskey & Cohen, 1989; Ratcliff, 1990; cf. Rosenberg & Sejnowski, 1986; Sejnowski & Rosenberg, 1988). Previous reports have shown qualitative tendencies of the models when using somewhat arbitrary stimulus representations. The present study reports quantitative fits of the models to data from a specific learning experiment, using stimulus representations independently derived from multi-dimensional scaling.

4.1. Human Memory

A simple experiment was designed to demonstrate the ability of humans to retain knowledge of previously learned associations while learning new associations, the ability of ALCOVE to model that fact, and the inability of (extended) backprop to show such memory. Figure 12 shows the abstract structure of the learning task. Like the previous experiment, the stimuli were rectangles that varied in their height with interior segments that varied in their lateral position, and the task was to learn one of two category labels for each stimulus. Only three levels on each dimension were used, the two extremes from the previous experiment plus the level halfway between. The assignment of labels to each exemplar is indicated in Figure 12 by

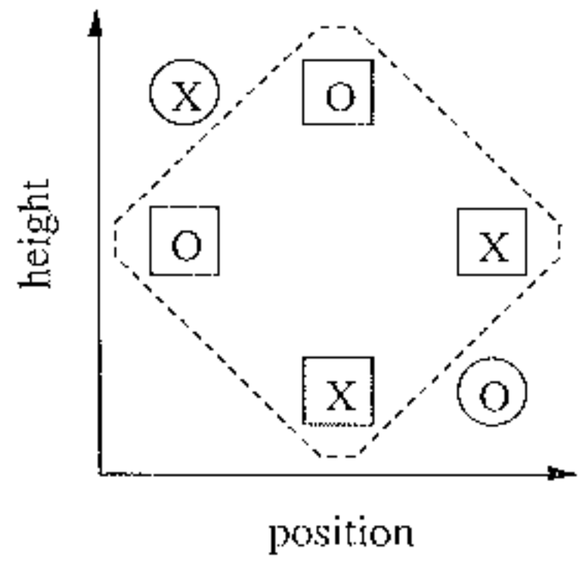


Figure 12. Category structure to demonstrate catastrophic forgetting in extended backprop. Category assignment is indicated by 'X' or 'O'. In the first phase of training, exemplars marked by squares (inside the dashed line) were shown with explicit category feedback, while exemplars marked by circles (outside the dashed line) were usually shown with only '?' as feedback. In the second phase of training, 'square' exemplars received '?' feedback, while 'circle' exemplars were shown with explicit feedback.

an 'X' or 'O'. Training consisted of two phases. In the first phase of training, exemplars marked by squares in Figure 12 (inside the dashed line) were shown with explicit category feedback, while exemplars marked with circles (outside the dashed line) were usually shown only '?' as feedback. In the second phase of training, the 'square' exemplars received '?' feedback, while the 'circle' exemplars were shown with explicit category labels. The transition between phases was not explicitly revealed to the subject.

It can be seen from Figure 12 that the 'square' exemplars constitute a condensation task. The category boundary for those exemplars is linear, along the right diagonal of the space, such that exemplars below the boundary are assigned to the 'X' category, and exemplars above the boundary are assigned to the 'O' category. The 'circle' exemplars can also be separated by a boundary along the right diagonal, but they require the opposite assignment to categories. Therefore, because (extended) backprop has receptive fields that cover half-spaces, with decision boundaries that are linear, it could show catastrophic forgetting of the 'square' exemplars when trained on 'circle' exemplars. On the other hand, ALCOVE uses hidden nodes with relatively localized receptive fields, and so it should show forgetting only to the extent that the receptive fields of the exemplar nodes overlap.

4.1.1. Procedure. Like the previous experiment, instructions were presented on the computer screen and read aloud by the experimenter. The instructions stated that the '. . . goal is to learn arbitrary labels for pictures'. The labels were the letters 'B' or 'N'. The instructions indicated that sometimes the computer would display the correct label, and sometimes only a '?'. The instructions continued: 'Each stimulus always has the same correct answer. For example, if on one trial the computer tells you that a certain stimulus was a 'B', then the correct answer for that stimulus is always 'B'—even if on some future trial the computer only displays a '?' for that stimulus. You should try to remember the correct label for each individual

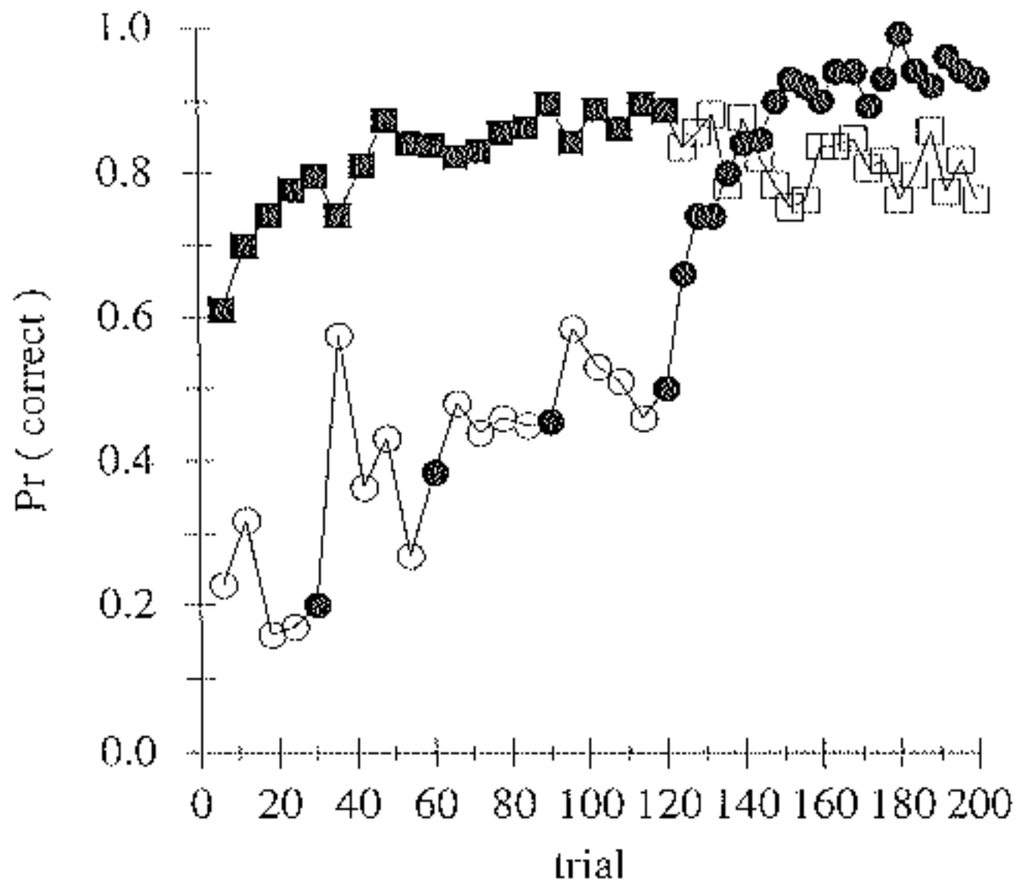


Figure 13. Human learning data for structure in Figure 12. Circles and squares show the mean proportion correct for 'circle' and 'square' exemplars in Figure 12. Filled markers denote trials on which there was explicit category feedback for those exemplars; open markers denote trials on which the feedback was a '?'.

stimulus.' Like the previous experiment, the instructions included two presentations of each of the training exemplars without feedback.

Training consisted of an uninterrupted sequence of 200 trials. Each subject saw the same fixed sequence of stimuli and feedback.

Phase 1: The first phase of training consisted of 20 sweeps through the six exemplars (120 trials), during which time the 'square' exemplars were always given explicit feedback ('B' or 'N'). The 'circle' exemplars usually got ambiguous feedback ('?') but were shown with explicit feedback every fifth block, so that subjects would realize that those exemplars really did have correct labels.

Phase 2: The second phase of training (continuous with the first and unannounced to the subject) consisted of 20 blocks of four trials in which the 'circle' exemplars were given explicit feedback and the 'square' exemplars were given ambiguous feedback. Every block of the second phase included the two 'circle' exemplars, and alternate blocks had the top-middle and bottom-middle exemplars (only) or the middle-left and middle-right exemplars (only), in order to keep the rate of ambiguous feedback down to 50%. Feedback was a single character ('B', 'N' or '?') displayed in the center of the screen for 3.0 s.

4.1.2. Subjects. A total of 48 subjects, recruited from the Indiana University campus area during Summer 1991, were paid $3.00 for their participation in the 30-minute experiment. Because of the relative paucity of subjects during the summer months, 28 of these subjects had previously participated in other category learning experiments (not reported in this article) using the same stimuli. This was not considered a problem for two reasons: first, at least 21 days had elapsed between experiments for all repeat subjects; and, second; the results from repeat subjects were qualitatively indistinguishable from the new subjects.

4.1.3. Results. Human learning data for this task are shown in Figure 13. Circles and squares show the mean proportion correct for 'circle' and 'square' exemplars in Figure 12. Filled markers denote trials on which there was explicit category feedback for those exemplars; open markers denote trials on which the feedback was a '?'. Two results should be noted. In the first phase, performance on exemplars with usually ambiguous feedback ('circles') noticeably improved after each presentation with explicit feedback, giving the learning curve a scalloped appearance. In the second phase, memory for the 'square' exemplars, now without explicit feedback, did not degrade dramatically, despite the fact that the 'circle' exemplars were learned very well during that time.

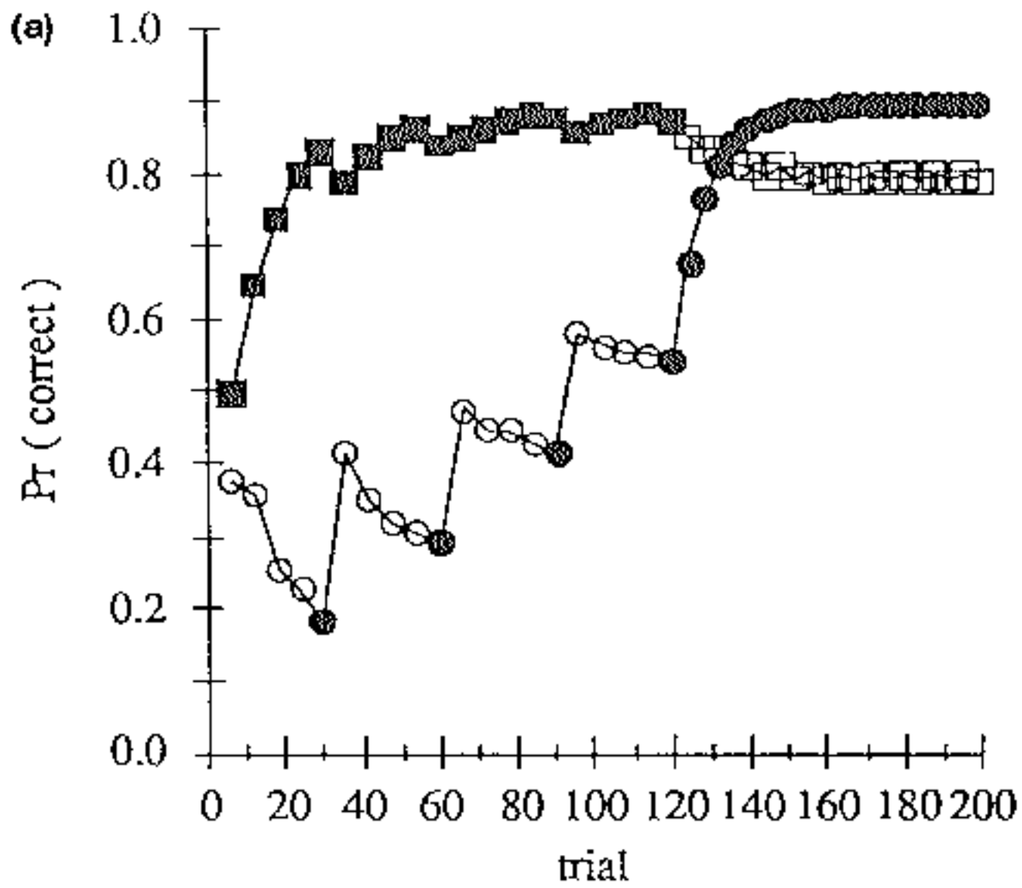
4.2. Fit of ALCOVE

ALCOVE was fitted to the data on a trial-by-trial basis, including the trials on which the feedback was only a '?'. Thus, the parameter values had to predict simultaneously learning (filled markers in Figure 13), generalization (open circles in Figure 13) and memory performance (open squares in Figure 13). The best fitting parameter values were $\phi = 1.08$, $\lambda_w = 0.282$, $c = 1.93$ and $\lambda_a = 0.00$, yielding an RMSD of 0.0727. The best-fitting predictions are shown in Figure 14. As expected, ALCOVE shows only slight forgetting during the second phase of learning. It also shows the scalloped appearance of the learning curve for the 'circle' exemplars during the first phase.

ALCOVE is able to remember the correct categorizations for the 'square' exemplars during the second phase because the receptive fields of the corresponding exemplar nodes overlap only a little with those of the 'circle' exemplars. It is important to realize, however, that some amount of overlap is crucial for capturing the human data. Without that overlap, completely isolated exemplar nodes would show zero forgetting during the second phase of learning, unlike humans, and isolated nodes would also show no generalization in the first phase, unlike humans. Some amount of overlap was also crucial for modeling filtration advantage, because isolated nodes could not affect attention strengths (in equation (8), either a_j^{hid} or $|h_{ji} - a_i^{\text{in}}|$ would be zero).

4.3. Fit of Backprop

In principle, (extended) backprop needs only four appropriately situated hidden nodes to classify accurately all six exemplars. (If the output nodes had threshold terms, or if the hidden nodes had activations in the range $[-1, +1]$ instead of $[0, +1]$, then only three hidden nodes would be needed.) Thus, the issue is not whether or not appropriate weights and thresholds *exist*, but whether or not backprop can *learn* them. In the simulations reported here, six randomly initialized hidden nodes were used, in order to match the number used in ALCOVE. The average performance of 200 random initializations was fitted to the data. The best fit of extended backprop yielded an RMSD of 0.208 (almost three times the RMSD achieved by ALCOVE) using parameter values of $\phi = 0.909$, $\lambda_{\text{out}} = 0.165$, $\lambda_a = 1.10$ and $g = 10.0$. The best fit is shown in Figure 14. Notice that there is catastrophic forgetting of the 'square' exemplars during the second phase (open squares), as performance drops to worse than chance level, unlike humans who maintained performance at about 80% correct. Another discrepancy between extended backprop and human performance is in the learning of the 'circle'



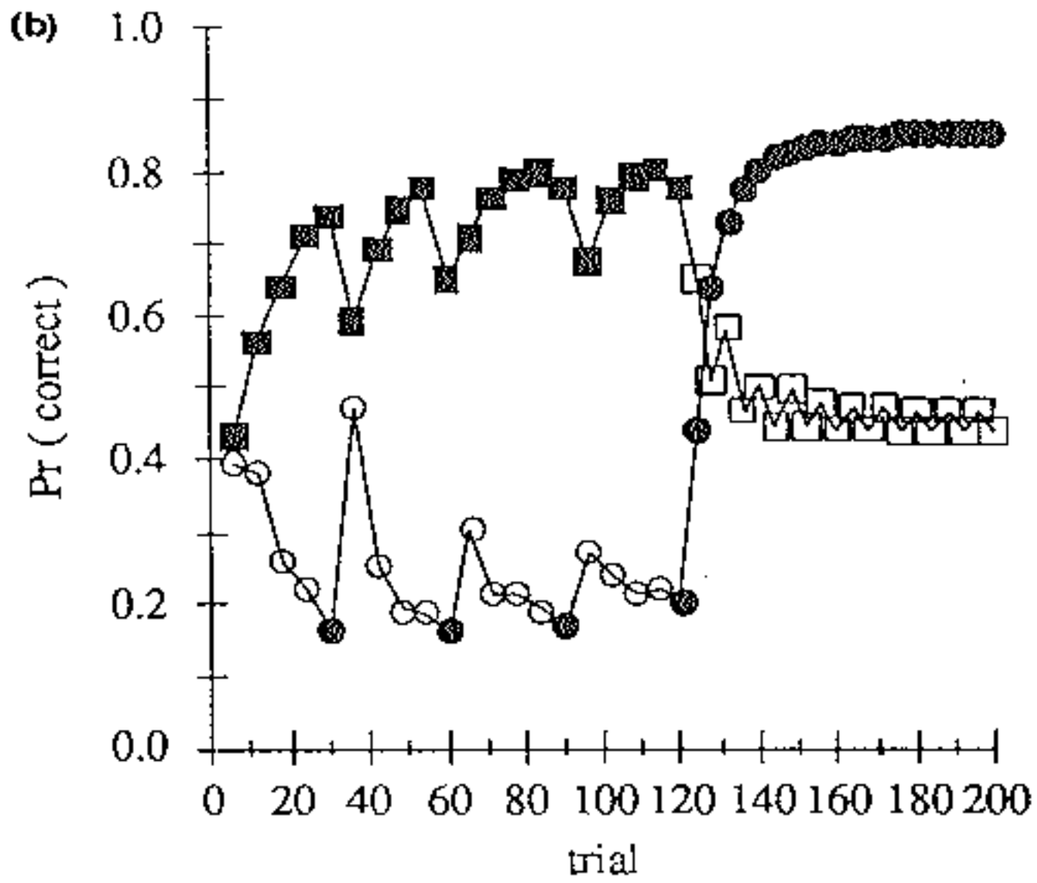


Figure 14. Best fit of (a) ALCOVE and (b) extended backprop to learning data in Figure 13.

exemplars during the first phase. Clearly, backprop cannot do as well as people, leaving performance at about 20% correct while people rose to about 50% correct.[5]

The cause of catastrophic forgetting in backprop is the huge receptive field of the hidden nodes. Figure 4 showed that a hidden node in backprop responds to an entire half-space of the stimulus space, whereas a hidden node in ALCOVE responds to a relatively localized region. Consequently, a given hidden node in backprop might be initially trained on stimuli lying in one region of input space, but subsequently be radically affected by new training stimuli in a vastly different region of input space. The extended version of backprop should suffer the same problem, because its hidden node receptive fields are unaffected by the inclusion

of dimensional attention strengths. In the present situation, the hidden nodes that learn first and fastest are those with receptive fields whose linear decision boundary lies along the right diagonal in Figure 12. In the first phase of training, such nodes learn that stimuli below the boundary are Xs, while those above the boundary are Os. Consequently, they show poor generalization during the first phase. In the second phase of training, the very same hidden nodes are optimally situated to respond to the 'circle' exemplars, only now the category mapping is reversed. The reversal causes catastrophic forgetting of the 'square' exemplars.

4.4. Discussion

Other researchers have demonstrated backprop's tendency to forget previously learned associations (McCloskey & Cohen, 1989; Ratcliff, 1990; cf. Rosenberg & Sejnowski, 1986; Sejnowski & Rosenberg, 1988), and I have previously demonstrated ALCOVE's resistance to catastrophic forgetting (Kruschke, 1992). What is new in the present study is the quantitative fit of the models to data from a specific learning experiment, using well-specified input and output representations. Also, my explanation of catastrophic forgetting in backprop emphasizes its over-sized hidden-node receptive fields (Figure 4), rather than the trajectory of connection weight learning in weight space emphasized by McCloskey and Cohen (1989; cf. Sejnowski & Rosenberg, 1988).

My claim here is not that backprop necessarily shows catastrophic forgetting in all situations; rather, that it is possible to construct at least some situations in which backprop must suffer that fate. Nor is it my claim that people never exhibit catastrophic forgetting, for surely there are situations in which they might. Nor is it my claim that ALCOVE cannot show catastrophic forgetting; it may if the phase 2 exemplars are sufficiently similar to the phase 1 exemplars, or if different dimensions are relevant in phase 2 than in phase 1. What has been demonstrated here is one situation in which people do not display catastrophic forgetting, nor does ALCOVE, but backprop must.

Some investigators might argue that there are techniques to prevent or significantly mitigate catastrophic forgetting in standard backprop (e.g. French, 1991; Hetherington, 1991; Lewandowsky, 1991; Sloman & Rumelhart, 1992). There are three reasons to doubt the sufficiency of those techniques to address all the concerns raised in this article, however. First, none of the techniques explicitly addresses the problem of backprop's inability to show filtration advantage, so that even if they can overcome catastrophic forgetting, they probably do not overcome the failure to properly selectively attend. Second, it is unclear if a combination of modifications for catastrophic forgetting and for filtration advantage would retain their desirable effects, or if they would interact and partially negate each other. Third, even if some combination of modifications was discovered to deal with filtration advantage and catastrophic forgetting, I believe another fundamental problem with backprop would remain: its oversensitivity to linear category boundaries. This is addressed in the next section.

5. Oversensitivity to Linear Boundaries

As discussed elsewhere (Gluck, 1991; Kruschke, 1990, 1992), backprop learns linearly separable categories faster than non-linearly separable ones, whereas people do not necessarily (Medin & Schwanenflugel, 1981). On the contrary,

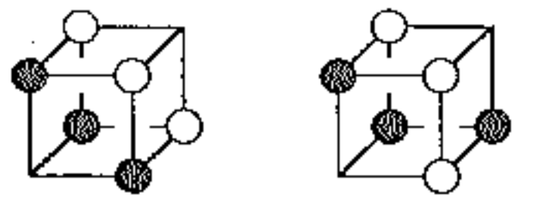


Figure 15. Category structures used by Medin and Schwanenflugel (1981, Experiment 4).

ALCOVE is only indirectly sensitive to the shape of category boundaries, and is primarily affected by the clustering of exemplars and their distribution over stimulus dimensions. In particular, whether a category boundary is linear or non-linear has no direct influence, and it is possible that non-linearly separable categories would be easier to learn than linearly separable ones.

A case in point comes from the work of Medin and Schwanenflugel (1981, Experiment 4). They compared learning rates for two category structures (Figure 15) such that one was linearly separable, whereas the other was not. The two structures were equated in terms of mean city-block distance between exemplars within categories and between exemplars from different categories. When human subjects were trained on the two structures, it was found that the linearly separable structure was no easier to learn than the non-linearly separable structure. This result is corroborated by earlier findings of Shepard *et al.* (1961). Although they were not concerned with linear separability *per se*, two of the category structures they studied can be construed as supersets of the ones studied by Medin and Schwanenflugel. Shepard *et al.*'s 'Type III' structure is non-linearly separable, and 'Type IV' is linearly separable. They can be visualized by adding filled dots to the lower left vertices of the cubes in Figure 15, and blank dots to the upper right vertices. For these larger structures, the mean city-block separations of exemplars within or between categories actually favors the linearly separable structure. Shepard *et al.*'s results showed no advantage for the linearly separable category; in fact, the non-linearly separable category was slightly easier to learn.

Other experimenters have found statistically reliable advantages for non-linearly separable categories, such as Medin and Schwanenflugel (1981, Experiment 3), Kemler Nelson (1984, Experiment 3 incidental condition) and Wattenmaker *et al.* (1986, Experiment 1 control condition and Experiment 2 no-theme condition). Nakamura (1985, no-theory condition) also found a trend toward faster learning of non-linearly separable categories.[6]

Backprop is unable to learn the non-linearly separable categories as fast or faster than the linearly separable ones in these situations (Gluck, 1991), unlike ALCOVE. Backprop's oversensitivity to linear boundaries is not a consequence of the learning algorithm, but is caused by the linear-sigmoid activation function used in the hidden nodes. Quite simply, the linear level-contours (see Figure 4) of the backprop hidden nodes can align—either via weight learning or by fortuitous initial orientation—with linear category boundaries, but cannot align (all in one piece) with non-linear boundaries.

The claim being made is that backprop cannot learn *these* particular non-linearly separable categories as *quickly* as these particular linearly separable ones, unlike humans (and unlike ALCOVE). There is no claim that backprop cannot learn non-linearly separable categories *eventually*.[7] Moreover, there is no claim that

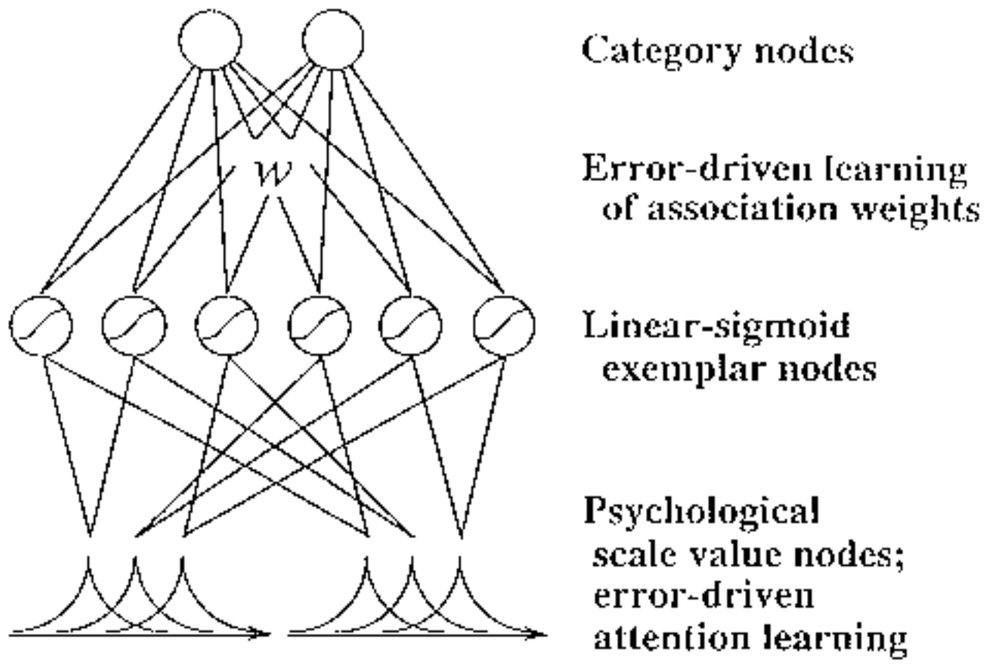


Figure 16. Structure of APPLE.

all linearly separable categories are easier for backprop to learn than all non-linearly separable categories—linearly separable categories for which all the exemplars are clustered tightly against the linear boundary might very well be harder for backprop to learn than non-linearly separable categories for which the exemplars from different categories are widely separated. Finally, there is no claim that *humans* learn all non-linearly separable categories faster than linearly separable ones.

I suspect that backprop's oversensitivity to linear boundaries cannot be overcome by modifying its learning procedure, because the origin of the problem is the form of the hidden node activation functions. The problem persists even when the hidden node weights are fixed at their random initial values—if a hidden node happens to align initially with the linear category boundary, it will learn the category distinction very quickly even without changing its incoming weights.

6. Modifying Backprop to Approximate ALCOVE

I have shown that backprop, using linear-sigmoid hidden nodes, suffers three significant problems as a model of human category learning. Is there no redemption for it? Are linear-sigmoid hidden nodes simply wrong for this application? No. In fact, one can construct an approximation to ALCOVE using linear-sigmoid hidden nodes and *place* coding of input dimensions (defined below), instead of the distance-similarity hidden nodes and amplitude encoding of input dimensions used in ALCOVE. For ease of reference, I will call this model APPLE (a homophone for the first syllables of 'APProximately ALcove'). Figure 16 shows its architecture. The input dimensions are each coded by an array of quasi-local nodes, positioned at all values that occur in the training set. For example, consider the exemplars from the catastrophic learning experiment (Figure 12). They take on three different heights and three different positions, so APPLE's implementation has three input nodes on each dimension, as illustrated in Figure 16. This type of encoding is sometimes called 'place' coding, as opposed to 'amplitude' coding used in ALCOVE (cf. Ballard, 1987; Hancock, 1989; Jeffress, 1948; Walters, 1987). Formally, the activation of an input node, centered on scale value h_i of dimension i, is given by

$$a^{\text{in}}_{h_i}(\psi_i) = \exp(-\alpha_i|h_i - \psi_i|) \tag{15}$$

where ψ_i is the psychological scale value of the stimulus.

Hidden nodes in APPLE correspond to training exemplars, as in ALCOVE. Each hidden node is linked to one (and only one) input node in each input dimension. Figure 16 shows six hidden nodes, corresponding to the six exemplars in Figure 12. Notice that each hidden node has only two incoming connections, one from each input dimension. Figure 16 shows each hidden node with a sigmoidal curve in it, indicating the sigmoidal activation function. Formally, the activation of hidden node j, linked to values h_i on dimensions i, is given by the linear-sigmoid function

$$a_j^{\text{hid}} = 1/\left(1 + \exp\left[-g\left(\sum_{h_i \in j} a_{h_i}^{\text{in}} - \theta_j\right)\right]\right) \tag{16}$$

where the sum is taken over all input nodes connected to hidden node j. It is convenient, though not essential, to set $\theta_j = N$, where N is the number of stimulus dimensions. In that case, the generalization gradient, as a function of distance in *psychological* space, closely resembles the exponential generalization gradient of a hidden node in ALCOVE, as shown in Figure 17. Approximations to *Gaussian* nodes using linear-sigmoid nodes have been previously noted by Hartman and Keeler (1991) and Mel (1990, Figure 5).

Learning in APPLE proceeds by gradient descent on sum-squared error, as in ALCOVE and backprop. The resulting learning rule for dimensional attention strengths is similar to the rule in ALCOVE:

$$\Delta\alpha_i = -\lambda_\alpha \sum_{h_i} \left[\sum_{\substack{\text{hid} \\ j}} \left(\sum_{\substack{\text{out} \\ k}} (t_k - a_k^{\text{out}}) w_{kj}^{\text{out}}\right) a_j^{\text{hid}} (1 - a_j^{\text{hid}}) g w_{jh_i}^{\text{hid}}\right] a_{h_i}^{\text{in}} |h_i - \psi_i| \tag{17}$$

where $w_{jh_i}^{\text{hid}}$ is 1 if h_i is connected to hidden node j, and is 0 otherwise. I have not simulated APPLE and fitted it to data because I do not expect it to be a dramatic improvement over ALCOVE (nor do I expect it to be dramatically worse than ALCOVE).

The point of describing APPLE is that what matters about ALCOVE is its use of error-driven learning, dimensional attention learning and quasi-local representation in psychological space (Kruschke, 1993). Those principles can be formalized in many ways; ALCOVE and APPLE are just two (Hurwitz, 1990, presents another). Alternative formalizations are not necessarily equivalent, however, because each suggests its own variations and extensions.

7. General Discussion

Standard backprop suffers three failures to model human category learning: in some cases when humans show filtration advantage, backprop will not; in some cases when humans retain memory for earlier phases of learning, backprop will not; and in some cases when humans learn non-linearly separable categories as fast as linearly separable ones, backprop will not. An alternative model, ALCOVE, does not suffer those failures, despite the facts that it, like backprop, is a feed-forward network that learns by gradient descent on error, and uses the same input and output representations as backprop in the simulations presented here. If the input dimensions are place coded with attention strengths, as in APPLE (Figure 16), instead of amplitude coded as in ALCOVE, then a network with linear-sigmoid hidden nodes can approximate the behavior of ALCOVE. It is not

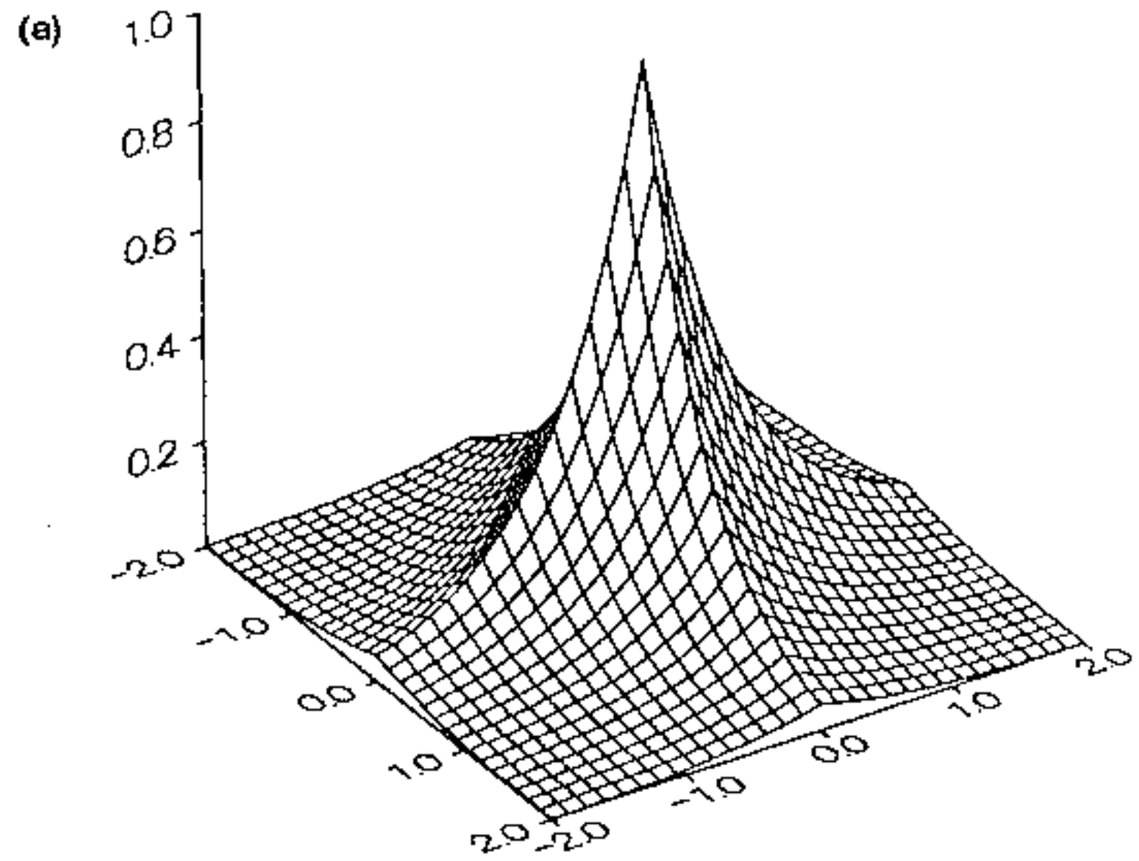


Figure 17. Activation profiles of hidden nodes in (a) ALCOVE and (b) APPLE, as a function of distance in psychological space. ((a) shows a_{ji}^{hid} as a function of ψ_1 and ψ_2 from equations (1) and (2), with $h_{j1} = h_{j2} = 0.0$, $\alpha_1 = \alpha_2 = 1.0$, $q = r = 1$ and $c = 1.5$; (b) shows a_j^{hid} as a function of ψ_1 and ψ_2 from equations (15) and (16), with $h_1 = h_2 = 0.0$, $\alpha_1 = \alpha_2 = 1.0$, $\theta_j = 2.0$ and $g = 3.5$.)

the learning mechanism, gradient descent on error, that causes problems for standard backprop. Rather, the failures are caused by the lack of dimensional attention strengths and the use of hidden nodes with non-local receptive fields in psychological space.

While the three problems addressed here have been qualitatively identified in one form or another by previous researchers, this article illustrated those problems with quantitative fits to human learning data from novel, honed-down experiments directly aimed at revealing the problems. Moreover, this article emphasizes that solving any one of the problems does not necessarily solve all three.

Does ALCOVE (or APPLE) solve all the problems of standard backprop? No. In particular, ALCOVE cannot learn as quickly as people do in some situations, such as the early trials of the filtration conditions (Figure 7). It is plausible that when confronted with such simple stimuli as used in these experiments (Figure 5), subjects will spontaneously hypothesize simple rules for explicit testing, such as 'tall rectangles are in category B', or 'if the segment is left of center, then it's in category N'. As those rules do solve the filtration problem, learning will be complete as soon as the correct rule is hypothesized. Like standard backprop, ALCOVE has no mechanism for explicit rule hypothesizing, hence situations for which rules play the dominant role might not be accommodated. Nevertheless, as I have suggested elsewhere (Kruschke, 1992), ALCOVE might interact with a rule-generating system, helping to steer its choice of rules by indicating with its dimensional attention strengths which dimensions are most relevant to the category distinction. The comments of McClelland and Jenkins (1991, p. 69), made with regard to their backprop model, also apply here: "Indeed, it must be acknowledged that there is a conscious, verbally accessible component to the problem solving activity that children and adults engage in when they confront a problem . . . . The model does not address this activity itself. However, it is tempting to suggest that the model captures the gradual acquisition mechanisms which establish the possible contents of the conscious processes."

Acknowledgements

Some of the research reported here was supported by Biomedical Research Support Grant RR 7031-25 from the National Institutes of Health, and by a Summer Faculty Fellowship from Indiana University. Thanks to Terry Bleizeffer, Deb Keller, Steve McKinley and Rita Randolph for running subjects. Thanks also to two anonymous reviewers of a previous version of this article. Portions of this research were reported at the Interfaces Conference on Categorization, Texas Tech University, Lubbock, TX, October 10–13, 1991; at the Twenty-Fourth Annual Mathematical Psychology Meetings, Indiana University, Bloomington, IN, August 10–13, 1991; and at the Thirteenth Annual Conference of the Cognitive Science Society, University of Chicago, Chicago, IL, August 7–10, 1991.

Notes

1. The scaling coordinates for the stimuli can be estimated to fit learning data, instead of independent similarity judgments, in which case those coordinates become additional free parameters in the learning model. This approach was explored by Nosofsky and Kruschke (1992).
2. The version of ALCOVE described here is *exemplar based*: a hidden node is placed at the location of each training exemplar prior to training. This is a reasonable approach for modeling the experiments presented here, because subjects were pre-exposed to the stimuli. Alternatively, hidden nodes could be scattered randomly across the stimulus space, forming a *covering map* of the space (Kruschke, 1990, 1992). The covering map does not store traces of training exemplars exclusively, but instead acts as a blanket of localized receptors, much like the coarse coding described by Hinton *et al.* (1986). This was the original conceptualization of ALCOVE (Kruschke, 1990), and ALCOVE stands for Attentional Learning COVEring map. The covering-map approach obviates prior knowledge of the training exemplars, but runs into the problem for simulation of an exponential explosion in the number of hidden nodes as the number of dimensions increases. A third approach is to start with zero hidden nodes and recruit new ones as novel stimuli are detected (Hurwitz, 1990). This requires additional mechanisms and parameters for novelty detection, but might prove to be the most general-purpose approach.

3. The inability of ALCOVE and backprop to learn the initial trials fast enough is addressed in the general discussion.
4. A related variation on back propagation was described by Mozer and Smolensky (1989), but in their model the hidden weights varied freely and the attention strengths did not change during learning, taking on only the values 1.0 or 0.0 as determined by post-training analysis.
5. I also fitted extended backprop using 12 hidden nodes. The quantitative fit was better, RMSD = 0.162, but the same qualitative results occurred. In particular, there was still severe forgetting, with performance dropping nearly 30 percentage points during the second phase.
6. The latter three papers emphasized the malleability of the advantage for non-linearly separable categories under the influences of semantic context and learning strategy.
7. It is well known that, given sufficient hidden nodes, there exist connection weights to approximate any given set of training pairs to any degree of accuracy (e.g. Hornik *et al.*, 1989). Whether such a set of weights can be learned from any arbitrary initial values is an unanswered question.

References

Anderson, J.R. (1991) The adaptive nature of human categorization. *Psychological Review*, **98**, 409–429.

Ballard, D.H. (1987) Interpolation coding: a representation for numbers in neural models. *Biological Cybernetics*, 57, 389–402.

Bower, G.H. & Hilgard, E.R. (1981) *Theories of Learning*, 5th edn. Englewood Cliffs, NJ: Prentice-Hall.

Broomhead, D.S. & Lowe, D. (1988) Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2, 321–355.

Cohen, J.D., Dunbar, K. & McClelland, J.L. (1990) On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological Review*, 97, 332–361.

Elman, J.L. (1989) *Representation and Structure in Connectionist Models*. Crl Technical Report 8903, Center for Research in Language, University of California at San Diego.

Estes, K.W. (1986) Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General*, 115, 155–174.

Estes, K.W. (1988) Toward a framework for combining connectionist and symbol-processing models. *Journal of Memory and Language*, 27, 196–212.

French, R.M. (1991) Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, pp. 173–178.

Garner, W.R. (1974) *The Processing of Information and Structure*. Hillsdale, NJ: Erlbaum.

Gluck, M.A. (1991) Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science*, 2, 50–55.

Gluck, M.A. & Bower, G.H. (1988) Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, 27, 166–195.

Gottwald, R.L. & Garner, W.R. (1972) Effects of focusing strategy on speeded classification with grouping, filtering and condensation tasks. *Perception & Psychophysics*, 11, 179–182.

Gottwald, R.L. & Garner, W.R. (1975) Filtering and condensation tasks with integral and separable dimensions. *Perception & Psychophysics*, 18, 26–28.

Grossberg, S. (1987) Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science*, 11, 23–64.

Hancock, P.J.B. (1989) Data representation in neural nets: an empirical study. In D. Touretzky, G. Hinton & T. Sejnowski (Eds) *Proceedings of the 1988 Connectionist Models Summer School*. San Mateo, CA: Morgan Kaufmann, pp. 11–20.

Hanson, S.J. & Burr, D.J. (1991) What connectionist models learn: learning and representation in connectionist networks. *Behavioral and Brain Sciences*, 13, 471–518.

Hanson, S.J. & Gluck, M.A. (1991) Spherical units as dynamic consequential regions: implications for attention, competition and categorization. In R. P. Lippmann, J. Moody & D. S. Touretzky (Eds) *Advances in Neural Information Processing Systems 3*. San Mateo, CA: Morgan Kaufmann.

Hartman, E. & Keeler, J.D. (1991) Predicting the future: advantages of semilocal units. *Neural Computation*, 3, 566–578.

Hetherington, P.A. (1991) *The Sequential Learning Problem in Connectionist Networks*. Master's thesis, McGill University, Montreal, Quebec, Canada.

Hinton, G.E., McClelland, J.L. & Rumelhart, D.E. (1986) Distributed representations. In R. E. Rumelhart & J. L. McClelland (Eds) *Parallel Distributed Processing*, Vol. 1, Chapter 3. Cambridge, MA: MIT Press, pp. 77–109.

Hornik, K., Stinchcombe, M. & White, H. (1989) Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359–366.

Hurwitz, J.B. (1990) *A Hidden-pattern Unit Network Model of Category Learning*. Ph.D. thesis, Harvard University.

Jeffress, L.A. (1948) A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, **41**, 35–39.

Kemler, D.G. & Smith, L.B. (1978) Is there a developmental trend from integrality to separability in perception? *Journal of Experimental Child Psychology*, **26**, 498–507.

Kemler Nelson, D.G. (1984) The effect of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior*, **23**, 734–759.

Kolen, J.F. & Pollack, J.B. (1990) Scenes from exclusive-or: back propagation is sensitive to initial conditions. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, pp. 868–875.

Kruschke, J.K. (1990) *A Connectionist Model of Category Learning*. Ph.D. thesis, University of California at Berkeley. Ann Arbor, MI: University Microfilms International.

Kruschke, J.K. (1992) ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22–44.

Kruschke, J.K. (1993) Three principles for models of category learning. In G. V. Nakamura, R. Taraban & D. L. Medin (Eds) *Categorization by Humans and Machines*. Special volume in *The Psychology of Learning and Motivation Series*, Vol. 29. San Diego: Academic Press, in press.

Kruschke, J.K. & Movellan, J.R. (1991) Benefits of gain: speeded learning and minimal hidden layers in back-propagation networks. *IEEE Transactions on Systems, Man and Cybernetics*, **21**, 273–280.

Kruskal, J.B. (1964) Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115–129.

Lehky, S.R. & Sejnowski, T.J. (1988) Network model of shape-from-shading: neural function arises from both receptive and projective fields. *Nature*, **333**, 452–454.

Lewandowsky, S. (1991) Gradual unlearning and catastrophic interference: a comparison of distributed architectures. In W. E. Hockley & S. Lewandowsky (Eds) *Relating Theory and Data: Essays on Human Memory in Honor of Bennet B. Murdock*. Hillsdale, NJ: Erlbaum.

Luce, R.D. (1963) Detection and recognition. In R. D. Luce, R. R. Bush & E. Galanter (Eds) *Handbook of Mathematical Psychology*. New York: Wiley, pp. 103–189.

McClelland, J.L. & Jenkins, E. (1991) Nature, nurture and connections: implications of connectionist models for cognitive development. In K. VanLehn (Ed.) *Architectures for Intelligence*. Hillsdale, NJ: Erlbaum.

McCloskey, M. & Cohen, N.J. (1989) Catastrophic interference in connectionist networks: the sequential learning problem. In G. Bower (Ed.) *The Psychology of Learning and Motivation*, Vol. 24. New York: Academic Press.

Medin, D.L. & Schaffer, M.M. (1978) Context theory of classification learning. *Psychological Review*, **85**, 207–238.

Medin, D.L. & Schwanenflugel, P.J. (1981) Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, **7**, 355–368.

Mel, B.W. (1990) *The Sigma–Pi Column: A Model of Associative Learning in Cerebral Neocortex*. CNS Memo 6, California Institute of Technology, Pasadena, CA.

Minsky, M.L. & Papert, S.A. (1969) *Perceptrons*. Cambridge, MA: MIT Press. (1988 expanded edition.)

Moody, J. & Darken, C.J. (1989) Fast learning in networks of locally-tuned processing units. *Neural Computation*, **1**, 281–294.

Mozer, M.C. & Smolensky, P. (1989) Skeletonization: a technique for trimming the fat from a network via relevance assessment. In D. S. Touretzky (Ed.) *Advances in Neural Information Processing Systems I*. San Mateo, CA: Morgan Kaufmann, pp. 107–115.

Nakamura, G.V. (1985) Knowledge-based classification of ill-defined categories. *Memory and Cognition*, **13**, 377–384.

Nosofsky, R.M. (1986) Attention, similarity and the identification–categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39–57.

Nosofsky, R.M. (1987) Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **13**, 87–108.

Nosofsky, R.M. & Kruschke, J.K. (1992) Investigations of an exemplar-based connectionist model of category learning. In D. L. Medin (Ed.) *The Psychology of Learning and Motivation*, Vol. 28. New York: Academic Press, pp. 207–250.

Pavel, M., Gluck, M.A. & Henkle, V. (1989) Constraints on adaptive networks for modeling human

generalization. In D. S. Touretzky (Ed.) *Advances in Neural Information Processing Systems 1*. San Mateo, CA: Morgan Kaufmann, pp. 2–10.

Poggio, T. & Girosi, F. (1990) Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, **247**, 978–982.

Posner, M.I. (1964) Information reduction in the analysis of sequential tasks. *Psychology Review*, 71, 491–504.

Ratcliff, R. (1990) Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review*, **97**, 285–308.

Rescorla, R.A. & Wagner, A.R. (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds) *Classical Conditioning: II. Current Research and Theory*. New York: Appleton-Century-Crofts.

Robinson, A.J., Niranjan, M. & Fallside, F. (1988) *Generalising the Nodes of the Error Propagation Network*. Technical Report CUES/F-INFENG/TR.25, Cambridge University Engineering Department, Cambridge, UK.

Rosenberg, C.R. (1987) Revealing the structure of NETtalk's internal representations. In *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, pp. 537–554.

Rosenberg, C.R. & Sejnowski, T.J. (1986) The spacing effect on NETtalk, a massively-parallel network. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, pp. 72–89.

Rosenblatt, F. (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, 386–408.

Rumelhart, D. & McClelland, J.L. (1986) PDP models and general issues in cognitive science. In D. E. Rumelhart & J. L. McClelland (Eds) *Parallel Distributed Processing*, Vol. 1, Chapter 4. Cambridge, MA: MIT Press, pp. 110–146.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986a) Learning internal representations by error propagation. In J. L. McClelland & D. E. Rumelhart (Eds) *Parallel Distributed Processing*, Vol. 1, Chapter 8. Cambridge, MA: MIT Press, pp. 318–362.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986b) Learning representations by back-propagating errors. *Nature*, **323**, 533–536.

Schneider, W. (1988) Micro Experimental Laboratory: an integrated system for IBM PC compatibles. *Behavior Research Methods, Instruments & Computers*, **20**, 206–217.

Seidenberg, M.S. & McClelland, J.L. (1989) A distributed, developmental model of word recognition and naming. *Psychological Review*, **96**, 523–568.

Sejnowski, T.J. & Rosenberg, C.R. (1988) Learning and representation in connectionist networks. In M. S. Gazzaniga (Ed.) *Perspectives in Memory Research*. Cambridge, MA: MIT Press., pp. 135–178.

Shanks, D.A. & Gluck, M.A. (1991) *Tests of an Adaptive Network Model for the Identification, Categorization, and Recognition of Continuous-dimension Stimuli*. Report 9103, Department of Cognitive Science, University of California, San Diego.

Shepard, R.N. (1957) Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. *Psychometrika*, **22**, 325–345.

Shepard, R.N. (1962) The analysis of proximities: multidimensional scaling with an unknown distance function, I and II. *Psychometrika*, 27, 125–140, 219–246.

Shepard, R.N. (1964) Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, 1, 54–87.

Shepard, R.N. (1987) Toward a universal law of generalization for psychological science. *Science*, 237, 1317–1323.

Shepard, R.N., Hovland, C.L. & Jenkins, H.M. (1961) Learning and memorization of classifications. *Psychological Monographs*, 75(13). Whole No. 517.

Sloman, S.A. & Rumelhart, D.E. (1992) Reducing interference in distributed memory through episodic gating. In A. F. Healy, S. M. Kosslyn & R. M. Shiffrin (Eds) *From Learning Theory to Cognitive Processes: Essays in Honor of William K. Estes*. Hillsdale, NJ: Erlbaum.

Stork, D.G. (1989) Is backpropagation biologically plausible? In *IJCNN International Joint Conference on Neural Networks*. New York: IEEE, pp. II-241–II-246.

Taraban, R., McDonald, J.L. & MacWhinney, B. (1989) Category learning in a connectionist model: learning to decline the German definite article. In R. Corrigan, F. Eckman & M. Noonan (Eds) *Linguistic Categorization*. Philadelphia: John Benjamins, pp. 163–193.

Walters, D. (1987) Properties of connectionist variable representations. In *Program of the Ninth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, pp. 265–273.

Wattenmaker, W.D., Dewey, G.I., Murphy, T.D. & Medin, D.L. (1986) Linear separability and

concept learning: context, relational properties, and concept naturalness. *Cognitive Psychology*, 18, 158–194.
Winer, B.J., Brown, D.R. & Michels, K.M. (1991) *Statistical Principles in Experimental Design*, 3rd edn. New York: McGraw-Hill.

Appendix: Scaling the Stimulus Space

Procedure

Subjects were shown written instructions on the computer screen, read aloud by the experimenter. The instructions included examples of all eight stimuli, each presented twice for 1.5 s in a random order, without any response from the subject. Subjects were told to rate similarities by pressing one of the keys from 1 to 9 (on the top row of the keyboard, not on the numeric keypad), where 1 meant 'least similar' and 9 meant 'most similar'. Subjects were encouraged to use the entire response scale, and were told that there was no emphasis on response speed and that there was no objectively correct answer.

Trials consisted of presenting two stimuli sequentially, for 1.70 s each, separated by a 0.70 s blank screen. A response prompting screen then appeared, which included a ruler-like scale from 1 to 9 with its ends marked 'least similar' and 'most similar', respectively.

Subjects saw each of the $\binom{8}{2} = 28$ pairs four times in each order, for a total of 224 trials. Each subject saw a different random sequence. There was a break in the middle of the sequence, when subjects could rest briefly.

A total of 50 subjects participated for partial credit in an introductory psychology course.

Results and Scaling Solution

The psychological coordinates of the stimuli were computed using similarity ratings averaged across subjects, repetitions and order of presentation, so that each datum was the mean of 400 observations. It was assumed that the psychological dimensions were independent, so that, for example, the psychological height of a rectangle did not depend on the position of the interior segment. Therefore, each psychological dimension had four values, corresponding to the four physical values.

Coordinates were selected to minimize the *stress* (Kruskal, 1964) between the city-block separations of pairs and the best *linear* prediction of separation on the basis of mean similarity rating. Stress was measured as $(\sum_i (d_i - \hat{d}_i)^2 / \sum_i d_i^2)^{1/2}$, where d_i is the city-block separation of pair i and $\hat{d}_i = a\delta_i + b$ is the linear prediction of distance on the basis of the observed similarity, δ_i. A linear regression function was used instead of a monotone regression function because the sparse distribution of stimuli could lead to degenerate solutions (Kruskal, 1964) when using a monotone function. Because the overall scale of the space was arbitrary, the psychological distance between the two shorter heights was fixed at 1.0. There were five remaining separations to be estimated, plus the two constants in the linear regression function, for a total of seven free parameters for fitting the similarity ratings.

The correspondence of city-block distances between pairs of stimuli in the psychological space, with rated similarities of those pairs, is shown in Figure 18. The fit had a stress of 0.06169, with distance accounting for more than 98% of the

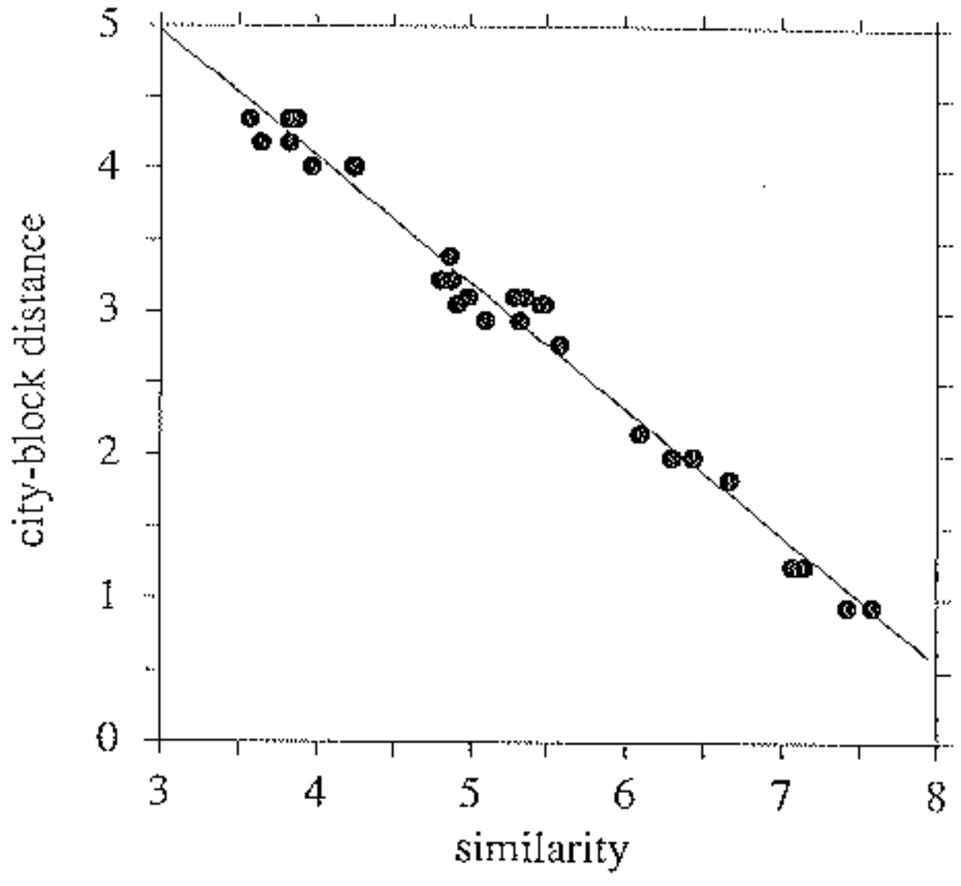


Figure 18. Correspondence of rated similarities and city-block separations for pairs of stimuli. Distance accounts for more than 98% of the variance in similarity ratings.

variance in similarity ratings. A Euclidean metric was also tried, and resulted in a minimal stress of 0.14567, much worse than the city-block metric. (When a monotone regression function was used, the minimal stress was zero because the sparse distribution of stimuli allowed degenerate solutions.) The best-fitting psychological coordinates of the stimuli are shown in Figure 8, in the main text of this article.