Representing Categories in Artificial Neural Networks Using Perceptually Derived Feature Networks

Robert B. Branstrom (branstrm@socrates.berkeley.edu)
Department of Psychology; 3210 Tolman Hall
University of California
Berkeley, CA 94720

Abstract

How might categories be represented in artificial neural networks while satisfying biological constraints? This article proposes using feature networks, an architecture based on two types of neural organization in perceptual systems, receptive fields and topographic representation. Using these two organizing principles, category features are represented in distributed networks that allow precise, graded or probabilistic interpretations. Simulations are illustrated that show these networks have characteristics consistent with human behaviors of assimilation, contrast, and chunking. A brief discussion and simulation show how these feature networks can be combined associatively to form complex multiple-feature categories. Implications of the architecture for representation and the nature of symbol processing are discussed.

Introduction

Regardless of the nature of the representation (i.e., visual image, verbal, etc.), categories are a foundational aspect of higher level cognition. The nature of categories remains a topic of considerable debate. The classical, or Aristotelian, view is that characteristics or traits define categories: things which have those characteristics are in the category and those which do not are not. This is simplistic, because sometimes something is in a category but does not have all necessary characteristics. For example, a three-legged animal that chases cats and cars would still be classified as a dog, even if it doesn't have the requisite four legs. Two approaches, both dealing with uncertain information, have evolved to address this problem. The first approach is probabilistic, asserting that something may be in a category if its characteristics are likely, rather than necessary. Thus the three-legged dog is still a dog because dogs usually, but not always, have four legs. The second approach applies the concept of graded structure (Rosch, 1973), asserting that membership in the category is a matter of degree, not an all-or-nothing feature. Thus a three-legged dog would still be a dog, albeit not as good an example as a four-legged dog. The condition "has four legs" is only partly satisfied, so the animal is not as good an example of a dog.

What approach might be taken to model categories? The classical view can be represented by formal set theory. Modifications to this view have been made to accommodate the probabilistic view and the graded structure view. In the probabilistic view, something is in the category if it has, say, eight of the necessary 10 conditions (Medin & Smith, 1984). The graded structure view has been approximated by fuzzy set theory (Zadeh, 1965). However, these views were developed for their formal properties, not their biological realism, so they don't offer plausible mechanisms that might underlie categorization processes.

An approach that steps closer to the biological structures of the brain is connectionism. Loosely, connectionist (or artificial neural network) models, assert that the brain is composed of many highly interconnected neurons, and that the processing power of the brain comes from these many connections. Network models of categorization typically represent categorical structure as a set of nodes representing characteristics (cf., Anderson, 1995). The characteristic may be absent or present (valued at 0 and 1, respectively). This vector of characteristics can also have graded values between 0 and 1. These values could represent either the probability or degree of the characteristic being present.

While these network models of categorization have useful functional characteristics, it's generally accepted that they still do not represent an approach that is close to the brain's actual organization. Among other things, real brains are expected to have more distributed representations for high level concepts. Anderson (1995, p. 345-6) proposed a number of principles to guide the development of "natural data representations," based on what is known about vertebrate nervous systems. These are worth summarizing here:

1. Similar events should give rise to similar representations.
2. Things should have separate representations if they need to be separated, thus categories could be separated by their features.
3. If something is important it should be represented by multiple elements.
4. Preprocess information as much as possible in the hardware.

5. Make the representation flexible so it is not problem specific.

Anderson also asserts (p. 346) that it would be easy to use "rather crude spatial means--say, spatially organized excitation and inhibition--to emphasize or deemphasize one or another aspect of the computation." Following Anderson's guidelines, this article proposes a network model of categorical and conceptual representation in which each feature is represented by a set of spatially organized nodes. The model accommodates both probabilistic and graded structure theories. The paper is organized as follows. First, two key structures of brain organization in perceptual systems are introduced and adapted for representation of category features. Then a number of simulations are provided to illustrate key behavioral characteristics of the features model. A proposal is then made for how these feature networks could be interconnected to provide an aggregate model of a category or concept. Finally, some implications of the model for cognitive science are discussed.

Representing single attributes

There are two common characteristics of perceptual systems that are spatially based. The first is the organization of sensory inputs using receptive fields. Receptive fields are sets of input cells that are interconnected such that closer cells have a common effect (excitatory or inhibitory) on the next level of processing. More distant cells have the opposite effect. In two dimensions, these are described as center-on, surround-off if the closer cells are excitatory, or center-off, surround-on if the closer cells are inhibitory. The second characteristic of perceptual systems is analogical representation of the physical world in neural structure. In visual and haptic systems this is spatially based topographic representation, and in the auditory system it is frequency based tonographic representation. In both cases, the principle is the same: values close to each other in the physical world are close to each other in the neural structure.

Sometimes these structures are combined, with rows of interconnected receptive fields. In the visual system, this architecture is responsible for the well-known effect of Mach bands, in which differences in contrast in input data are enhanced at edges to increase contrast sensitivity. This effect is illustrated in Figure 1. The lower graph of the figure shows the specific inputs to each cell. The upper graph shows the output pattern across many cells, including the enhanced contrast where the input pattern changes.

This model applies the same architecture (rows of interconnected receptive fields) to features of categories and concepts. Two important observations are important here. First, features are usually scalar in nature, i.e., they carry ordinal (and sometimes higher level) information.

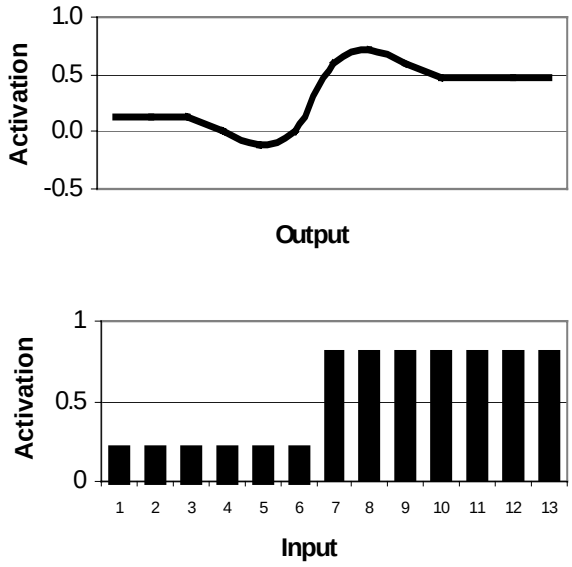

Figure 1: Edge contrast--Interconnected receptive fields enhance differences in input values at the edge where the difference occurs.

For example, dogs typically have fur. This can be represented on a scale from no-fur (Mexican hairless) to heavily furred (St. Bernard). Second, characteristics may be precise (24 inches tall) or vague (about 24 inches tall). This model allows for both of these characteristics. The ordered nature of a feature (i.e., the degree to which it holds) is mapped topographically onto the ordered organization of the nodes in the network. Precise values are represented as single nodes and vague values are represented as a cluster of adjacent nodes.

Several comments are in order before describing the model more specifically. First, the use of conceptual topographic mappings (as compared to physical or spatial topographic mappings) shouldn't be surprising if we take seriously the claim of evolutionary biologists, who argue that the easiest way to create a new structure is to borrow an old one. Second, representations of number are assumed to be at the level of an interval scale, so that both the order and distance between nodes is relevant to the representation. Third, nodes in the model's feature network are not suggested to be at the level of neurons, nor are they intended to be physically adjacent to each other. The *organization* of the nodes is the important factor; if this architecture holds in real brains it is expected that each node would be made up of many neurons and that connections would be distributed over wide areas. Finally, it should be noted that the idea of distributing features over multiple nodes was used by Shultz and Lepper (1996) to model cognitive dissonance. They distributed features across two-node polarized pairs.

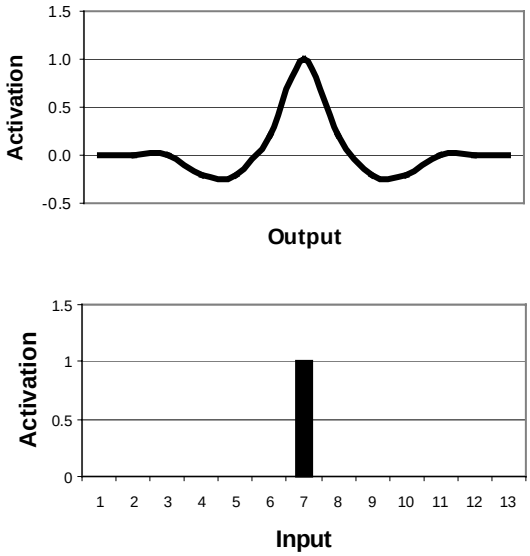

Figure 2: Point-valued representation--Input to a single node results in a characteristic "Mexican hat" output pattern.

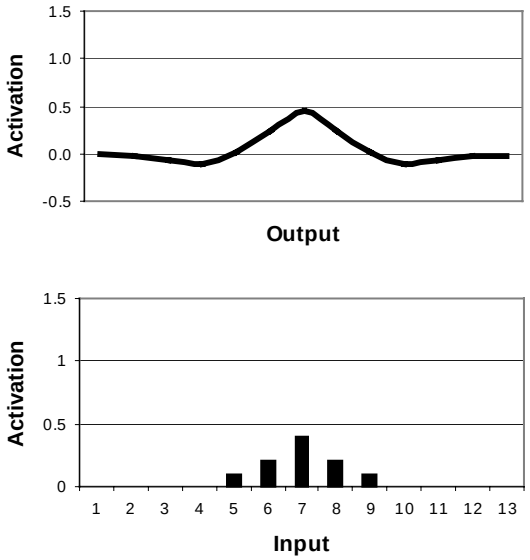

Figure 3: Vague-valued representation--Input to several adjacent nodes results in the same output pattern, but one that is more dispersed.

The model

The model was created on a spreadsheet. Specifically, a one-dimensional row of nodes (cells) was used to represent a feature. Each node's activation was calculated as the sum of its input and the weighted-sum of the inputs of the six nearest nodes. Neighboring node inputs were all weighted at 0.2 of their actual value, and were positive for adjacent nodes and negative otherwise. In other words, the neural representation was a set of one-dimensional, overlapping, center-on/surround-off receptive fields. Inputs are modeled as values from 0 to 1, and outputs can be either positive or negative. (Although this latter effect is neurally unrealistic--neurons don't have negative activations--it is assumed this is reasonable given the usual positive base activation rate, which may be reduced. The zero base rate is used for simplicity of exposition.)

Point-valued vs. vague-valued representations

Representations may be either point-valued or vague. This is modeled as either a single input or input spread across several nodes. Figure 2 shows a point-valued representation and Figure 3 shows a vague-valued representation. In both cases, the effects are similar: from the center of input the activation spreads slightly to neighboring cells, with closer cells being less activated than the central point and further cells being inhibited to negative values.

Vague representations may be interpreted as either probabilistic or graded. Thus, in Figure 3, the input value for 7 may be interpreted as a 40% probability of 7 occurring or as 7 to degree 0.4. When interpreted as probabilities, it isn't required that these values sum to 1. This is consistent with empirical findings on subjective estimates of probabilities (Edwards, 1961).

Assimilation and contrast effects

In addition to probabilities, judgments of similarity are also subjective. Sherif, Taub, and Hovland (1958) found that, when comparing two weights, subjects' estimates of the weight of one item depended on the similarity of the comparison weight. When the two weights were very similar, subjects shifted their weight judgments of the test weight (relative to when there was no comparison weight) towards the value of the comparison weight. This effect (or bias) they labeled assimilation. As the difference between weights increased, subjects shifted their estimates of the test weight more than the actual changes. This effect (or bias) was labeled contrast. In short, when two items were compared, the subjective judgment of difference depended upon the amount of the actual difference. Small initial differences were reduced so the two items appeared more similar than they actually were, while larger initial differences were enhanced so the two items appeared more different than they actually were.

The feature model yields the same effects. Figure 4 illustrates two point-valued inputs that are close to each other, yet still separated by another node. Their output, however, is merged into a single lump. (In this case, the output is two-peaked. The actual shape depends upon several factors, including the number of cells between inputs, the size of the receptive fields and the value used to weight neighboring cell inputs.)

A contrast effect, which occurs when the distance between the initial inputs is increased, is illustrated in Figure 5. The contrast occurs in two ways. First, the

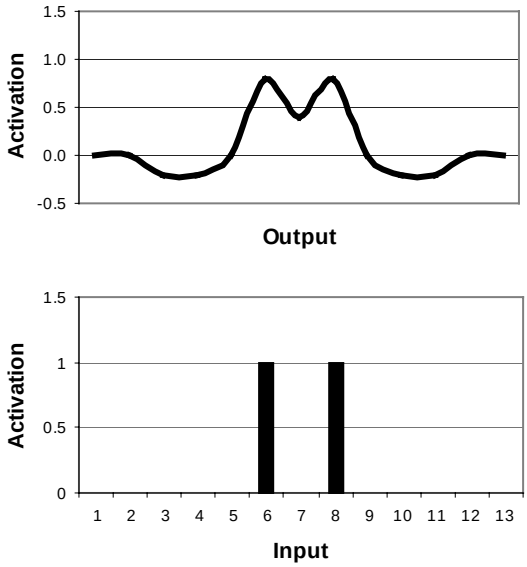

Figure 4: Assimilation effect--When two inputs are close to each other, the outputs from the feature network are merged into a single output.

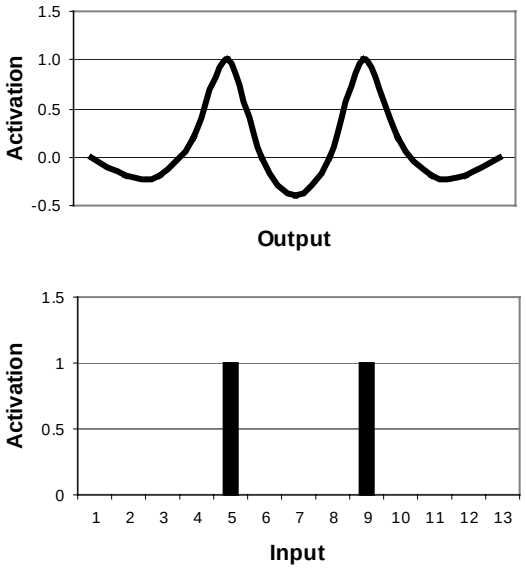

Figure 5: Contrast effects--When two inputs are slightly further apart, the outputs enhance the difference, both horizontally and vertically

activation value of the most intermediate node between the input values is inhibited below its normal base rate of zero, heightening the vertical contrast with the activation values of the nodes where the input actually occurs. Second, the "center of mass" of the output activations is shifted horizontally, slightly away from the actual value where the input occurs. This is seen in the actual output activation values. For example, input occurs at node 5, which has the highest output (activation value equal to 1). But node 4's activation is .2 and node 6's activation is 0. This asymmetry, in effect, shifts the mean activation of the representation for that input slightly away from its actual value.

Several observations are in order here. First, the effects are a result of the size of the receptive field. The assimilation effect occurs when the center (excitatory) parts of the receptive fields overlap and the contrast effect occurs when the surround (inhibitory) parts of the receptive fields overlap. Second, the assimilation effect could put a lower bound on what differences can be perceived; in effect they represent a just noticeable difference (Gregory, 1987, p. 405) for whatever is represented in the network. Third, if learning features from environmental inputs has created appropriately sized receptive fields, these effects are functionally adaptive. Essentially, assimilation allows for very small (and likely irrelevant) differences to be ignored, because they are merged and treated as one. Slightly larger (and likely more important) differences, which might not otherwise be noticeable, have their differences enhanced. (Even larger differences, which presumably would be easier to notice, aren't enhanced at all because the receptive fields of nodes receiving inputs don't overlap at all.) Fourth, these effects occur with vague representations as well as the illustrated point-valued representations. Finally, this contrast effect is similar to the peak shift found in stimulus learning (Hanson, 1959). Peak shifts occur when a correctly learned stimulus (which generalizes over a symmetric gradient) must be discriminated from a new, closely related stimulus. The original stimulus gradient shifts slightly, creating an asymmetric gradient, but one that enhances discrimination. Because peak shifts are learned, they occur over time, whereas contrast effects occur immediately in real time. But both are adaptive mechanisms that enhance contrast.

Chunking

One of the best known effects in cognitive science is chunking, the combination of several smaller bits of information into a single larger piece (Miller, 1956). When multiple pieces of information are represented as inputs in the feature network, assimilation and contrast effects provide a type of chunking. Figure 6 illustrates this, with seven inputs in two clusters of five and two, separated by one node with no input. The resulting output is two distinct "chunks," which could be called "low" and "high" on the particular feature in question.

Multiple features

Typically, categories are made up of items with complex combinations of multiple features. This section begins an exploration of this issue by considering how feature networks might be combined to represent more complex concepts and categories. Due to the dynamic complexities of interconnected features, this section provides only a sketch of how multiple attributes might be represented.

Because each feature is represented as a network of

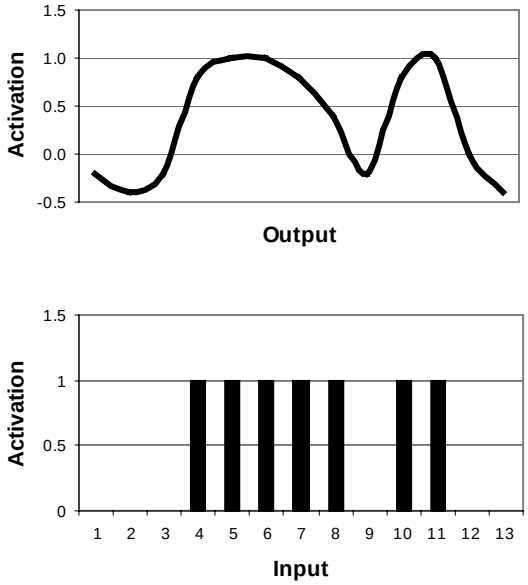

Figure 6: Chunking--When many input nodes are activated, their outputs are clustered into related groupings, such as "high" and "low."

ordered values, these networks can be related based on correlations between features. For example, the ability to fly is correlated with the presence of wings. Both are characteristics of birds and other flying species. Thus positive connections can be made between corresponding nodes in the two different attributes, such as good flying ability and good wings. Similarly, negative correlations can be made between opposite ends of the network. Figure 7 illustrates how these connections would be made from two nodes in a "flying ability" feature network to two nodes in a "wings" feature network. The straight-across connections are positive (shown as solid lines) representing positive correlations, and the diagonal connections are negative (shown as dashed lines), representing negative correlations. The double arrows on all connections represent that the connections are bilateral, that is they are mutually excitatory or inhibitory. This allows a dynamic interplay between the features, such that each node includes among its inputs the activations of the other feature's nodes from the previous iteration. These recurrent connections require a more complex formulation of the node activation functions, particularly the use of decay to dampen each node's activations over time. In the simulations presented here, correlative connections were weighted ± 0.2 and each node's activation value was decayed 80% from the prior period before computing the net input values.

When interconnected in this way, activation spreads from one feature to another. Figure 8 shows the spread of activation from an activated feature (flies well) in one period to a secondary feature (has good wings) in the following period. Two interesting characteristic of the

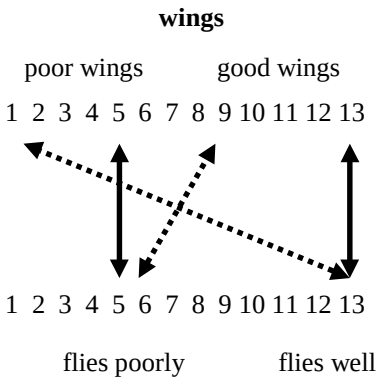

Figure 7: Multiple feature networks can be connected based on the correlations of features. Solid lines represent positive correlations and dashed lines represent negative correlations. Here, good flying ability (node 13) is positively correlated with good wings (node 13) and negatively correlated with poor wings (node 1). Double arrowheads indicate that the connections are bilateral.

secondary feature's output are the weaker level of activation relative to the activated feature and the drop in activation on the poor side of the scale, creating a contrast with the activated end of the feature network. The first characteristic is due to the weight of the correlation connections being less than one. The second characteristic results from the inhibitory connections that cross over to the opposite end of the secondary feature. The net effect of these two characteristics is that the activation level is lowered, but this is offset by an induced contrast effect.

Categories

Treating categories as features can extend the use of interconnected feature networks to categories. For example, "birdness" is descriptive of a category, but can also be treated as a feature that is correlated with features like flight, wings, feathers, and egg laying. Because they are correlated, all the features of the category would be connected to the category network. Thus, networks for features like flying ability, wings, feathers, lays eggs, etc. would all connect to a bird feature network. When some of the features of being a bird are activated, the activation spreads to other features, including the bird feature.

Levels of categories (superordinate, basic, and subordinate) also appear to be easily computed in this structure, because the assimilation and contrast effects of the feature networks allow for generalization to higher category levels via chunking, and discrimination between lower level categories via contrast effects. Further simulations are needed to explore these dynamics.

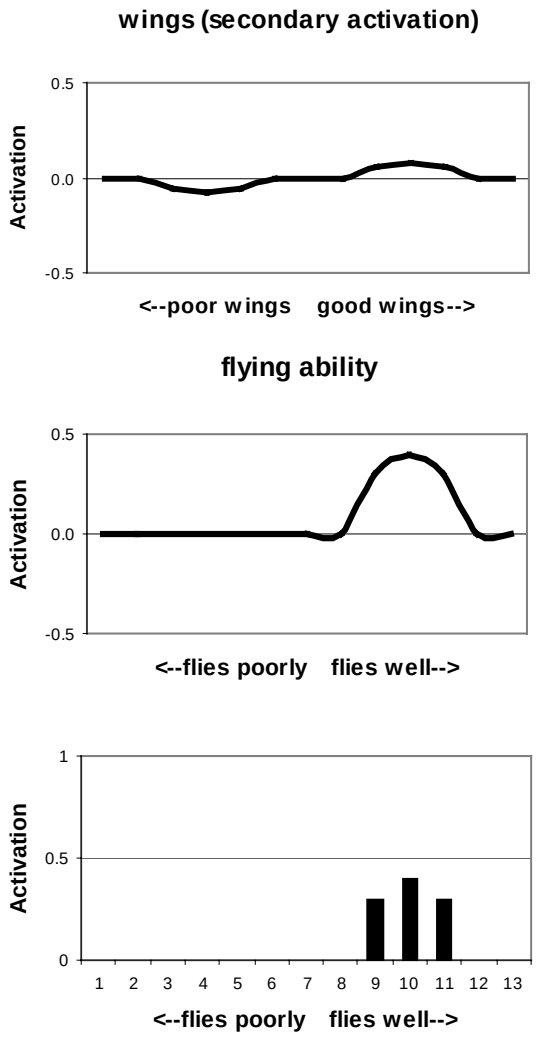

Figure 8: Activation of a correlated feature--Input (bottom graph) characterized vaguely as "good flyer" causes a similarly vague output of the "flying ability" feature network (middle graph). Both positive and negative correlations with the "wings" feature network (top graph) result in a slight contrast effect.

Conclusions

If categories provide the foundation of higher level thought, then their representation in neural structures is an important nut to crack. The proposed method of representing categories via their features in ordered feature networks is promising because it is simple and based on known patterns of neural organization. These networks allow for crisp, vague, and probabilistic representations. Perhaps most unusual, they provide a natural way to dynamically generalize and bifurcate concepts because of their assimilation and contrast effects. While further research about the characteristics of these networks (especially more complex interconnected feature networks) is needed, their ability to perform these basic tasks is intriguing.

Because these networks provide a means of representing symbolic information, they may shed light on the nature of symbolic thought. Those who view the mind as a symbolic processor and those who view the mind as a vast connectionist network have reached an uneasy truce. While not held universally, the view promoted by Smolensky (1988) is common: The mind is a symbol processor that runs on top of a neural network computing platform. The feature network model presented here suggests that this simple dichotomy may be unrealistic because the nature of the symbol processing itself may be important. In particular, dynamic grouping and splitting of fuzzy neural representations (i.e., generalizing and discriminating) and associations between correlated features may characterize thought more than logical operations.

Acknowledgements

I appreciate the comments provided by Christine Diehl, Janek Nelson, and Michael Ranney on an earlier version of this paper.

References

Anderson, J. A. (1995). *An introduction to neural networks*. Cambridge, MA: MIT Press.

Edwards, W. (1961). Behavioral decision theory. *Annual Review of Psychology, 12*, 473-498.

Gregory, R. L. (Ed.). (1987). *The Oxford companion to the mind*. New York: Oxford Press.

Hanson, H. M. (1959). Effects of discrimination training on stimulus generalization. *Journal of Experimental Psychology, 58*, 51-65.

Medin, D. L., & Smith, E. E. (1984). Concepts and concept formation. *Annual Review of Psychology, 35*, 113-138.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for information processing. *Psychological Review, 63*, 81-97.

Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories. In T. Moore (Ed.), *Cognitive Development and the Acquisition of Language*: Academic Press.

Sherif, M., Taub, D., & Hovland, C. I. (1958). Assimilation and contrast effects of anchoring stimuli on judgments. *Journal of Experimental Psychology, 55*, 150-155.

Shultz, T. R., & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review, 103*(2), 219-240.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*, 1-74.

Zadeh, L. (1965). Fuzzy sets. *Information and Control, 8*(3), 338-353.