A Computational Model which Learns to Selectively Attend in Category Learning

Lingyun Zhang Garrison W. Cottrell
lingyun,gary@cs.ucsd.edu
UCSD Computer Science and Engineering
9500 Gilman Dr., La Jolla, CA 92093-0114 USA

***Abstract*— Shepard et al. made empirical and theoretical investigation of the difficulties of different kinds of classifications using both learning and memory tasks [Shepard et al., 1961]. As the difficulty rank mirrors the number of feature dimensions relevant to the category, later researchers took it as evidence that category learning includes learning how to selectively attend to only useful features, i.e. learning to optimally allocate the attention to those dimensions relative to the category [Rosch and Mervis, 1975]. We built a recurrent neural network model that sequentially attended to individual features. Only one feature is explicitly available at one time (as in Rehder and Hoffman's eye tracking settings[Rehder and Hoffman, 2003]) and previous information is represented implicitly in the network. The probabilities of eye movement from one feature to the next is kept as a fixation transition table. The fixations started randomly without much bias on any particular feature or any movement. The network learned the relevant feature(s) and did the classification by sequentially attending to these features. The rank of the learning time qualitatively matched the difficulty of the categories.**

***Index Terms*— category learning, selective attention, sequential processing, recurrent network, fixation transition table**

I. Introduction

Shepard et al. (1961) examined how people learn simple concepts. The concept was a classification task. The stimuli were 8 objects which differed in 3 binary features (Figure 1). 4 of the 8 stimuli are assigned to class one and the remaining 4 are assigned to class two. There are 70 possible ways to assign the classes (combinations of four out of eight). The 70 possible categories belong to 6 basic types (Figure 2):

- Category type I: categories in which the classification is based on the value of only one feature (the example in Figure 2 is based on color: black ones are in class 1 and white ones are in class 2.)
- Category type II: categories in which the classification is based on the values of two features (the example in Figure 2 is based on color and shape: black triangles and white squares are in class 1 while white triangles and black squares are in class 2)
- Category type III, IV and V: categories in which classification is based on all three features, but some of the individual feature or feature pairs give information about the classification (this will be further discussed when examining the mutual information between the features and categories)
  - Type III: only two features are needed per instance (it can be solved by a height 2 decision tree).
  - Type IV: each feature is useful alone (classification based on any single feature can achieve 75% accuracy.)
  - Type V: only one feature is useful by itself.
- Category type VI: categories in which classification is based on all three features and no individual features or feature pairs give any information about the classification. This category type includes only the two possible 3d XOR classifications.

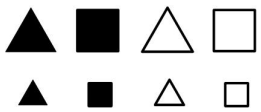

Fig. 1. 8 stimuli differed in 3 binary features: shape (circle or triangle), size (big or small) and color (black or white) (adapted from [Shepard et al., 1961])

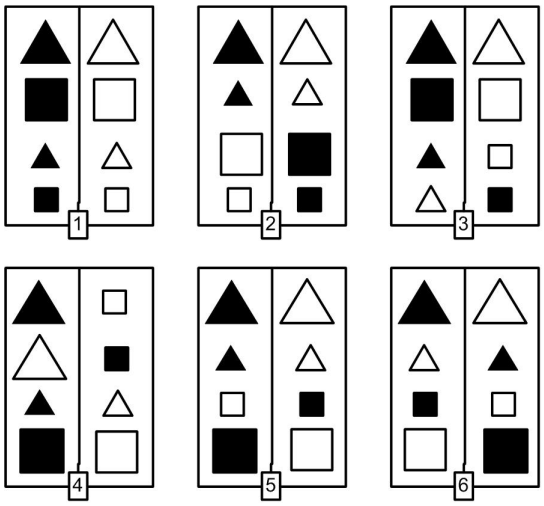

Fig. 2. 6 basic categories (adapted from [Shepard et al., 1961])

Shepard et al. found that the difficulty ordering among these basic category types is type I<II<(III, IV, V)<VI. When the features are separated in space, the difficulties increase but the ranking remains the same. Since the ordering suggests that the difficulty goes up with the number of feature dimensions

that the category type based on, it was taken as evidence that category learning includes learning how to selectively attend to only the feature dimensions useful for classification [Rosch and Mervis, 1975].

Recent work by Rehder and Hoffman(2003) further investigated selective attention in category learning by separating the feature dimensions in space and looking at the eye movement data by eye tracking. This is theoretically interesting because by using eye tracking, what feature dimensions have been actually attended to and when people attend to them can be explicitly measured. This tells us more than just learning speed. The 3 binary feature dimensions were realized by 3 paris of text symbols ($ or $\bar{c}$, ? or ! and + or -). An example of the stimuli is shown in Figure 3. They had experiments on category type I, II, IV and VI, given type III, IV and V have similar difficulties. The subjects were instructed to guess the class of a presented stimulus by pressing one of two buttons and they got feedback immediately after the guess. The 8 different stimuli were presented randomly in blocks of 8. The learning continued for 28 blocks or ended early if the subject guessed correctly for 4 blocks in a row. The eye tracker was set up to record subjects' eye movement during the learning. They replicated the category difficulty ordering: $one < two < four < six$. Figure 4 shows the average number of features fixated for learners in each category type in each block. For subjects who ended early in learning, the data for the last block was used for remaining blocks. It shows that toward the end of the learning, the category one group only examined about 1 feature, the category two group only examined about 2 features and the category four and six groups usually examined all 3 features. The authors suggested that this confirmed that subjects allocated their attention (as measured by eye movements) to only those features needed to solve the classification problem. They discussed two categorization theories based on computational models: a connectionist model, ALCOVE [Kruschke, 1992], which predicts that attention weights gradually shift toward better performance as learning experiences accumulate, and a rule based model RULEX [Nosofsky et al., 1994], which predicts that learners will start from only one single feature. ALCOVE explains the human data better given that the subjects mostly start by attending to all three features, then stop attending to the unrelated ones. However, ALCOVE does not predict the sudden shift to one feature in type I which is followed by error free performance. Also, ALCOVE processed all the features in parallel. In Rehder and Hoffman's experiments, subjects can only attend to features sequentially because no two features can be foveated at same time. Note that this is only true for the stimuli whose features are separated in space (Figure 3) but not for the stimuli whose features are not (Figure 1.) The constraint that subjects can only sequentially attend to each feature could be the reason that category learning is harder when the features are spatially separated, though the difficulty rank does not change. In this paper, we built a recurrent network which attended to features sequentially. It learned the relevant features while learning the category. Section two explains our model; section three reports the simulation results; section four examines the mutual information between feature(s) and the category; section five concludes with some discussion.

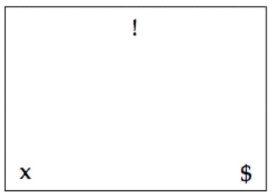

Fig. 3. The stimuli used by Rehder and Hoffman. Note that the features are spatially separated. (from [Rehder and Hoffman, 2003])

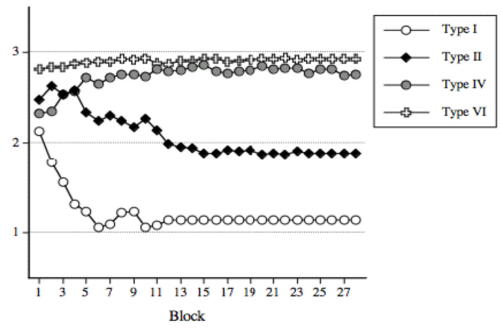

Fig. 4. The number of features fixated for each category learning. This is an average. A single subject shows a sudden change for type I. (from [Rehder and Hoffman, 2003])

II. Recurrent Model Which Learns the Relevant Features and the Category

In this section, we will describe our model which simulates the concept learning process. The model is inspired by Rogers and Casteren's recurrent network model of active vision for object recognition which could classify simple patterns in a positional invariant way and combine spatial features with dynamic fixations [Rogers and Casteren, 2003].

Our model consists of a classifier of a recurrent two layer back propagation neural network and a fixation transition table (FTT) whose entries are the probabilities of the next fixation given the current fixation (Figure 5). The classifier is a combination of both Elman's network and Jordan's network in the sense that both the output and the hidden layer are forwarded to the next step as input. Each binary feature is presented as a binary variable taking a value of 0 or 1 (i.e. ?=0, !=1) and so is the class label. The output consists of two units, which stand for the two categories. The input includes the current fixation, the feature at this fixation, and the output and hidden layer activations from last step.

The FTT contains the conditional probabilities of fixation transitions, which are updated according to the performance of the classifier. It probabilistically decides the next fixation upon the current one. Thus the eye movement decision making in the model is a Markov decision process with a table look

up stochastic policy. The policy is learned in a matter similar to temporal difference learning.

When the next fixation is decided, both the new fixation and the feature at this fixation are provided to the classifier's input. The output would then give a category confidence. The confidence could be low because of a lack of information (i.e. not every relevant feature has been examined yet). The output layer activation and hidden layer activation are copied to the input of the next step (the initialization of these at the first step is discussed later in this section.) Note that at every step, only the feature that is currently fixated is explicitly present in the input, which simulates the fact that subjects can attend to only one feature at a time since the features are separated in space. Information from earlier steps is only implicitly contained in the hidden layer, which can be taken as a dynamic working memory.

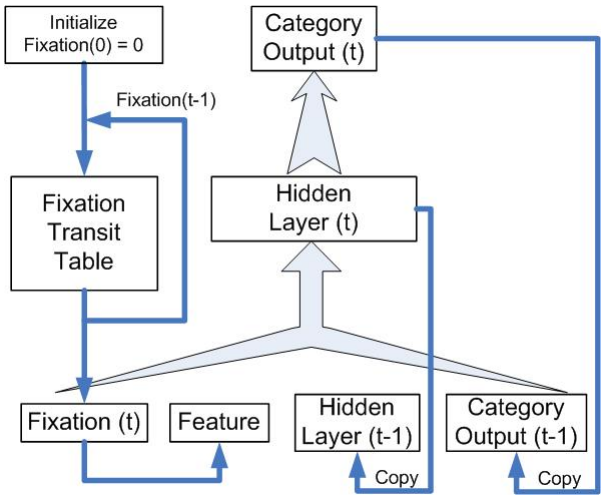

Fig. 5. Framework of our model.

The activation function of the hidden layer is a scaled tanh function [LeCun et al., 1998]:

$$f(x) = 1.7159\ tanh(\frac{2}{3}x) \tag{1}$$

30 hidden units are used in the results reported in the paper. The activation function of the output is softmax:

$$y_i = e^{a_i} / \sum_k e^{a_k} \tag{2}$$

Cross-entropy is used as the error function for categorization.

Each entry in the FTT is the conditional probability of the next fixation given the current fixation. Each column corresponds to a current fixation. Fixation 0 denotes the beginning of a trial (no current fixation yet). The first fixation is probabilistically decided by its transition probability to each feature. Fixations 1 to 3 denote the current fixation of features 1 to 3 respectively. Each row corresponds to a feature as the next fixation. The FTT is initialized before the learning process. Each entry is independently drawn from the uniform distribution in the range $[3, 4]$. Each column is then normalized to sum to 1. We initialize the FTT this way because we assume that there is no strong bias toward any particular feature at the very beginning, yet differences exist among individuals. Table I shows an example of an initialized FTT.

TABLE I
THE FIXATION TRANSIT TABLE, AN EXAMPLE OF INITIALIZATION

	Current Fixation			
Next Fixation	fix 0	fix 1	fix 2	fix 3
fix 1	0.3662	0.3129	0.3357	0.3174
fix 2	0.2995	0.3493	0.2932	0.3332
fix 3	0.3343	0.3377	0.3711	0.3494

The 8 patterns (000-111) were presented in random order within each block of 8, as in Rehder and Hoffman's experiments. For each pattern, the first fixation is generated by the first column of the FTT (conditional probability at the beginning of guessing a new pattern). The feature is set accordingly at the input. The category is set to [0.5 0.5] at the input as the prior of the two categories (this information is presented to subjects in Rehder and Hoffman's experiments). The hidden layer copy part of the input is set to zeros, simulating the fact that no information has been collected about the pattern at the very beginning. The information is then fed forward to the output layer. The next fixation is generated based on the current fixation and the corresponding column of the FTT. Then the corresponding feature is presented. The next step is carried on by feeding forward the new fixation, feature, last hidden layer activation and the category estimation at the output layer.

For each pattern, the network makes at most 3 fixations before making the final guess of the class according to the category output activation. The category with larger activation at the output layer after the last fixation is assigned to the pattern. If the confidence of a category is over 90% (the activation > 0.9), then the classifier stops early here and assigns the category to the pattern. After the guess, the correct answer is presented and the network knows whether it made a correct guess. This is similar Rehder and Hoffman's experiments, in which the human subjects would not have the feedback until they make the guess. The error is then back propagated to the classifier with standard back propagation for each step. The learning rate is α and there is a degrading ratio of θ each step which simulates the degrading of memory over time. Here we have an implicit assumption that fixations closer to the decision are remembered better, and so implicitly affect future decisions more. The degradation is inspired by the temporal difference learning of TD-Gammon (Gerald Tesauro 1995) which learns how to play the game of Backgammon.

In the process of making guess for each pattern, a fixation updating table (FUT), which has similar entries to the FTT, is initialized to have zero entries and accumulate during the fixations. At each new fixation, the FUT is degraded by a factor of λ and the corresponding entry of the eye movement in the updating table is incremented by γ.

$$FUT(t) \ = \ \lambda * FUT(t-1) \tag{3}$$

$$FUT(t)_{fix(t-1),fix(t)} = FUT(t)_{fix(t-1),fix(t)} + \gamma \quad (4)$$

At the end of each pattern guess, if the guess is correct, we take the sequence of movement as a good one. The FUT is then weighted by the confidence and added to the FTT to reinforce these movements by increasing the probabilities of the corresponding entries . The weighted FUT is subtracted from the FTT if otherwise.

$$FTT = \begin{cases} FTT + (y_i - 0.5) * FUT, & \text{correct} \\ FTT - (y_i - 0.5) * FUT, & \text{wrong} \end{cases} \quad (5)$$

y_i is the activation of the winner at the output layer.

Negative entries in FTT are set to 0 and those larger than 1 are set to 1. Then the FTT is normalized so that every column sums to 1. The network keeps learning through trials of blocks until it correctly classifies all the patterns for 50 blocks in a row or arrives an upper limit of blocks.

III. Results

We trained the network on basic category types I, II, IV and VI. The parameters used for the results reported here are showed in Table II. These parameters were decided empirically.

TABLE II

The parameters for learning the categories

α	θ	λ	γ
0.01	0.3	0.8	0.05

In Rehder and Hoffman's experiments the learning stops after 28 blocks (in each block the 8 patterns were shown once each in a random order) if not stopped earlier by making a correct guess for 4 blocks in a row. We set a maximum block number of $2,000$. The learning stops here, unless it ends earlier by classifying correctly 50 blocks in a row. Note that there is extremely low probability of classifying everything correctly for 50 blocks in a row (which contain 400 patterns) within $2,000$ blocks when the network only learns 7 out of 8 patterns. Thus when there are 50 blocks in a row correctly classified, we are confident that the network has learned the category of all the patterns.

Table III shows the results of learning and Table IV shows the human data from Rehder's experiments for comparison. One hundred networks were trained for each category. The portion learned is how many out of a hundred stopped early before reaching the maximum trial number. "Average blocks" is the average number of blocks being considered by those that stopped early (or those that learned before the maximum trial). The order of the learning time is qualitatively consistent with the difficulty of the categories.

Figure 6 shows the portion of each feature being fixated in the process of training. As in Rehder and Hoffman's experiments (Figure 4), the last block's data is used for blocks afterwards. Consistent with human data, during the learning of category one, the one feature that is relevant gets more and more fixations until the category is learned. Similarly, in the learning of category two, the two relevant features get more and more fixations. In the learning of categories four and six, the three features get about the same portion of fixations throughout the learning.

TABLE III

Learning for category type I, II, IV and VI

Category	1	2	4	6
Networks learned out of 100	100 100%	100 100%	96 96%	91 91%
Average blocks (Std)	126 (32)	246 (59)	446 (194)	1108 (382)

TABLE IV

The human data from Rehder's experiments [Rehder and Hoffman, 2003]

Category	1	2	4	6
Individuals learned out of 18	18 100%	18 100%	15 83%	10 56%
Average blocks	7.11	14.11	18.11	22.94

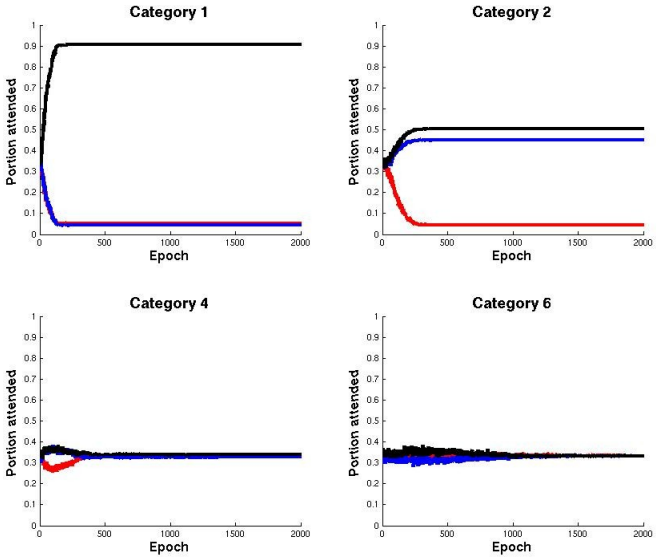

Fig. 6. The portion of each feature being fixated in category learning.

Table V shows some FTT exemplars of each category after learning. We observed some characteristics of the learned FTT and corresponding model behavior for each category:

- For category 1, a high value is always assigned to entry (1,0) after learning which means the network almost always starts from feature one, which is the only feature relevant to the category. The confidence of classification usually arrives 90% at the first step and the categorization ends here. So in most trials, the network ends up with only one fixation at feature one.
- For category 2, a high value is always assigned to either entry (1,0) or entry (2,0) which means a given network

almost always starts from either feature one or feature two which are the only features relevant to the category. If entry (1,0) is big, entry (2,1) is always big while entry (1,1) and entry (3,1) is always small. This makes sure that if the network starts from feature one, then the next feature being attended is feature two. Similarly, if entry (2,0) is big, entry (1,2) is always big while entry entry (2,2) and entry (3,2) are always small. In both cases, the FTTs make sure that features one and two will be attended in a certain sequence in the first two fixations. That is, for all input patterns, the converged sequence is either $1 \rightarrow 2 \rightarrow ?$ or $2 \rightarrow 1 \rightarrow ?$.

- For categories 4 and 6, similar to category 2, the FTT always converges to a certain fixation pattern, while the pattern is always a sequence of all three features being fixated. For examples shown in Table V, a pattern of $3 \rightarrow 2 \rightarrow 1$ for category 4 and a pattern of $1 \rightarrow 3 \rightarrow 2$ for category 6.

TABLE V

EXAMPLES OF FTT AFTER CATEGORY LEARNING

	Category 1			
	0	1	2	3
1	0.9956	0.4899	0.4164	0.3285
2	0.0042	0.1607	0.3750	0.3990
3	0.0002	0.3494	0.2087	0.2725
	Category 2			
	0	1	2	3
1	0.0011	0.0005	0.9991	0.2989
2	0.9987	0.7318	0.0005	0.3012
3	0.0002	0.2678	0.0004	0.3999
	Category 4			
	0	1	2	3
1	0.0009	0.6781	0.9999	0.0010
2	0.0000	0.3142	0.0000	0.9990
3	0.9991	0.0077	0.0001	0.0000
	Category 6			
	0	1	2	3
1	0.9988	0.0001	0.4282	0.0000
2	0.0006	0.0014	0.0000	1.0000
3	0.0006	0.9985	0.5718	0.0000

IV. MUTUAL INFORMATION BETWEEN FEATURES AND CATEGORIES

To further investigate how much information each feature conveys about a category quantitatively, we compute the mutual information between features and categories.

$$I(F;C) = H(C) - H(C|F) \tag{6}$$

In the equation, I denotes mutual information, C denotes category, F denotes feature and H denotes entropy. Entropy measures the uncertainty. Mutual information I(F;C) thus measures how much the uncertainty about the category is reduced by knowing the feature.

Table VI shows an example of each basic category type. The first column shows all the possible patterns (corresponding to the 8 stimuli) where x,y and z are the three features that each can be 0 or 1. Columns 2-7 are examples of a basic category type I-VI respectively. Patterns with labels 0 belongs to one class and those with 1 belongs to the other.

TABLE VI

BASIC CATEGORY TYPES IN BINARY

Patterns	Category					
xyz	I	II	III	IV	V	VI
000	0	1	0	1	0	1
001	0	1	0	1	0	0
010	0	0	1	0	1	0
011	0	0	1	1	1	1
100	1	0	1	0	1	0
101	1	0	0	1	0	1
110	1	1	1	0	0	1
111	1	1	0	0	1	0

Mutual information is then computed by the following equation and the results are shown in Table VII. Similarly we can compute the mutual information between pairs of features and the category. The results are shown in Table VIII.

$$I(F;C) = \sum_f \sum_c p(f,c) \log \frac{p(f,c)}{p(f)p(c)} \tag{7}$$

TABLE VII

THE MUTUAL INFORMATION BETWEEN INDIVIDUAL FEATURES AND CATEGORIES

	I	II	III	IV	V	VI
x	1	0	0	0.1887	0	0
y	0	0	0.1887	0.1887	0.1887	0
z	0	0	0.1887	0.1887	0	0

TABLE VIII

THE MUTUAL INFORMATION BETWEEN PAIRS OF FEATURES AND CATEGORIES

	I	II	III	IV	V	VI
xy	1	1	0.5	0.5	0.5	0
yz	0	0	0.5	0.5	0.5	0
zx	1	0	0.5	0.5	0	0

Note that if the mutual information is 0, it means just by this feature (or pair of features) we can do no better than random. While if the mutual information is 1, it means by this feature (or pair of features) we can classify perfectly. If the mutual information is somewhere between 0 and 1, we can do better than chance but not perfectly. Table VII shows that by attending to only one feature, we can classify type I perfectly, get some information about type III, IV and V, but cannot get any information for types II or VI. Table VIII shows that by attending to only two features, we can classify type II perfectly, get more information for type III, IV and V than by attending to one feature, but still cannot get any information for type VI. This may explain why category VI is more difficult than category III, IV and V although all of these categories involve three features. The reason could be that

category III, IV and V can be partially classified by attending to one or two features but category VI needs all three features to be considered to be better than random.

V. Discussion

Our model successfully learned the selective attention in the category learning task. It was simple and did not have an explicit mechanism to select features. It started from attending to all the features more or less equally, but ended up attending to relevant features much more than others through learning.

The mutual information between features and categories suggests that by greedily incrementing informative features according to mutual information would fail at least for type II, because the mutual information between individual features and the category are all zero for type II. The eye movement pattern of the human subjects suggests that instead of finding informative features one after another, it is more likely that all available information is considered at the beginning and unrelated features are later discarded in this category learning. This is consistent with our model and ALCOVE. The key difference between our model and ALCOVE is that ALCOVE is exemplar based and achieves selective attention by decreased attention weight on irrelevant features, while our model does so by fixating them rarely or not at all. It has been hypothesized that exemplar based classification processes are most likely to operate in domains involving integral dimension as opposed to separable dimension stimuli [Nosofsky and Palmeri, 1997]. Integral dimension stimuli are tend to be perceived and represented as unitary whole, while highly separable dimension stimuli may require serial processing or limited-capacity parallel processing [Garner, 1974], [Lockhead, 1972], [Shepard, 1964], [Shepard and Chang, 1963], [Treisman and Gelade, 1980]. We suggest that ALCOVE accounts for the process of selectively allocating attention in category learning when the features are not spatially separated, while our model accounts for the process when they have to be sequentially examined.

As discussed earlier in the result section, for categories two, four and six, where more than one feature is relevant to the category, our model always converges to a certain fixation pattern. That is, for all the eight stimuli, it examines the relevant features in the same sequence. This scan sequence is different from trial to trial due to different initialization of the FTT and the connecting weights of the network. Our model predicts that people ultimately converge on a certain fixation pattern during the process of category learning when the features are spatially separated. The fixation sequence of different patterns is consistent within an individual when the category is learned, while differences across individuals are expected.

Our model is very simplified in the sense that there are only 3 features to choose from and each feature is a binary variable. Also, the eye movements only depend probabilistically on the position of the last fixation but not on the feature values. We believe that sequential processing is what ultimately goes on in high levels of vision because of the limit of computing ability - that is why we have saccades. The interesting questions here are how the eye movements are directed and how the information gained across saccades is integrated. We would like to take this model as a preliminary work for a more realistic eye movement model which would learn complicated tasks such as face identification.

Acknowledgement

We thank Timothy T. Rogers for inspiring the idea of the model and valuable discussions, Jonathan Nelson, Eric Wiewiora and Hector Jasso for suggestions, Nicholas Butko and Eric Wiewiora for proof reading and Gary's Unbelievable Research Unit (GURU) for comments. This research was supported by NIMH grant MH57075 to GWC.

References

[Garner, 1974] Garner, W. (1974). The processing of information and structure. New York: Wiley.

[Kruschke, 1992] Kruschke, J. (1992). Alcove. *Psychological Review*, 99:22–44.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Orr, G. B., and Müller, K.-R. (1998). Efficient backprop. In *Neural Networks—Tricks of the Trade, Springer Lecture Notes in Computer Sciences*, volume 1524, pages 5–50.

[Lockhead, 1972] Lockhead, G. (1972). Processing dimensional stimuli: A note. *Psychological Review*, 79:410–419.

[Nosofsky and Palmeri, 1997] Nosofsky, R. and Palmeri, T. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, 104(2):266–300.

[Nosofsky et al., 1994] Nosofsky, R., Palmeri, T., and McKinley, S. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, 101:53–79.

[Rehder and Hoffman, 2003] Rehder, B. and Hoffman, A. (2003). Eyetracking and selective attention in category learning. In Alterman, R. and Kirsh, D., editors, *Proceedings of the 25th Annual Cognitive Science Conference*, Boston, Massachusetts, USA. Cognitive Science Society.

[Rogers and Casteren, 2003] Rogers, T. and Casteren, M. (2003). A simple recurrent network model of active vision for object recognition. Cognitive Neural Science Society. Note: presented as a poster.

[Rosch and Mervis, 1975] Rosch, E. and Mervis, C. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.

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

[Shepard and Chang, 1963] Shepard, R. N. and Chang, J. (1963). Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology*, 65:94–102.

[Shepard et al., 1961] Shepard, R. N., Hovland, C. I., and Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, 75(517).

[Treisman and Gelade, 1980] Treisman, A. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136.