Predicting the Ease of Human Category Learning Using Radial Basis Function Networks

Brett D. Roads
Department of Computer Science and Institute of Cognitive Science
University of Colorado Boulder
Boulder, CO 80309-0430
brett.roads@colorado.edu
present address:
Department of Experimental Psychology
University College London
26 Bedford Way
WC1H 0AP London UK
b.roads@ucl.ac.uk

Michael C. Mozer
Department of Computer Science and Institute of Cognitive Science
University of Colorado Boulder
Boulder, CO 80309-0430
mozer@colorado.edu

January 25, 2019

Abstract

A common hurdle in category learning research is determining the ease of learning a particular item or category. For example, when exploring different training policies it is desirable to know which items or categories are most difficult. Unfortunately, existing methods for determining ease-of-learning are costly for real-world stimulus sets. This work introduces a novel approach for computing ease-of-learning, which we refer to as ease values, by determining an embedding of the domain items and using simple category learning models to predict human performance. Instead of using behavioral data from human training experiments, our approach uses human similarity judgments to fit the free parameters of a radial basis function network, which serves as a simple category learning model. Results from two validation experiments demonstrate the efficacy of the approach and the superiority of a simple exemplar-based model over two alternative models. There are three advantages of the proposed approach. First, collecting similarity judgments is relatively cheap compared to hiring experts or conducting training experiments. Second, the approach is designed to work for arbitrary, real-world, visual domains. Third, the approach is relatively forgiving if ease values must be determined for a new expanded set of stimuli.

Visual categorization is a critical skill in many professions, including radiology, dermatology and satellite imagery analysis. The economic importance of visual categorization has motivated substantial research aimed at reducing the cost of training visual experts. One approach for improving training is to predict where learners are likely to make errors and adjust the training protocol appropriately. For example, a training protocol could introduce easy stimuli first (e.g., Hornsby & Love, 2014; McLaren & Suret, 2000; Roads, Xu, Robinson, & Tanaka, 2018). Accurate assessment of a learner's knowledge also requires knowing something about stimulus difficulty: if an assessment is composed solely of easy items, it will be challenging to distinguish between trained and untrained individuals. In this work, the relative ease of learning a particular exemplar is referred to as the *exemplar ease value* (EEV). The average EEV of all exemplars in a category is referred to as the *category ease value* (CEV). The objective of this work is to provide a cost-effective method for predicting EEVs and CEVs, generically referred to as EVs.

A variety of methods have been proposed for estimating EVs. In the machine learning literature, the likelihood of classification error is related to the distance between a stimulus and the category boundary. For example, in the case of learning a linear (hyperplane) separator, examples near the boundary require more

precise specification of the separator, and are therefore less likely to be classified correctly given limited training data. Machine learning approaches rely on a known representation in order to compute distance to the category boundary. In the human literature, analogous experiments have been done using synthetic stimuli whose representations are known in advance and whose distance from the category boundary can be determined (e.g., Spiering & Ashby, 2008). When a stimulus representation is unknown–as is typically the case when using real-world stimuli–the human literature employs three strategies. First, the distance from a category boundary can be empirically determined by collecting error statistics from humans (novices or experts) during the categorization task. The second approach is to norm stimuli to a given concept. For example, Salmon, McMullen, and Filliter (2010) collected ratings for a set of images in which participants rated each stimulus on its "graspability". These norms can then been used by other researchers to create arbitrary category boundaries (Khan, Mutlu, & Zhu, 2011; Lindsey, Mozer, Huggins, & Pashler, 2013). Third, expert-based norms can be obtained by hiring experts to rate the difficulty of each stimulus (e.g., Evered, Walker, Watt, & Perham, 2014). However, there are two problems associated with the existing approaches. First, it is costly to collect error statistics (via training experiments) or hire experts. Second, even when a representation is known, ease of learning may depend on factors other than distance to the category boundary.

The focus of our work is on leveraging human similarity judgments to reduce the overall burden of predicting EVs. We explore three predictive models. The first two models use human similarity judgments to determine a latent embedding (psychological representation) of the stimuli and then characterize human learning in terms of this embedding. One model is based on exemplars, the other on prototypes; both are variants of a radial basis function network. The third model is neutral to psychological theory and merely counts the frequency of different types of similarity judgments in order to determine EVs. Each model is evaluated by comparing the predicted EVs to empirically-derived EVs.

1 A cost-effective method for estimating ease values

Theories of human category learning provide an alternative approach for estimating EVs. If a model is capable of predicting the likelihood that a particular stimulus will be categorized correctly (e.g., Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky, 1986), then the model's predictions can be used to estimate EVs. For example, assuming the appropriate free parameters are already known, a category learning model could be used as a surrogate subject to collect error statistics. Alternatively, a category learning model that has successfully

mastered a task could be treated as a surrogate expert and used to estimate expert ratings.

The predictive power of a category learning model comes at a cost. Nearly every human category learning model relies on behavioral data to tune the model's free parameters. In a typical setting, category learning models are trained using behavioral data is collected from a human training experiment. For the reasons already discussed, we would like to avoid the costs associated with running a training experiment.

As an alternative, we propose an approach that fits a model's free parameters using human similarity judgments. Human simialrity judgments are less time-consuming to collect than running training experiments. We consider a class of simple category learning models based on Radial Basis Function Networks (RBFN). The proposed approach leverages the fact that the activations of an RBFN are modulated by free parameters that can be learned from human similarity judgments or fixed based on psychological theory. In the following three sections, we outline a simple class of RBFNs, detail how the free parameters are inferred, and describe two specific RBFN implementations.

1.1 RBFN category learning model

At its simplest, a RBFN consists of three layers: an input layer, a hidden layer, and an output layer. A number of free parameters govern how activation propagates through the network. These free parameters belong to one of three network components: the stimulus representation, the psychological similarity kernel, and the association weight matrix. These three components are implemented by commonly used category learning models, such as the Generalized Context Model (Nosofsky, 1986) and ALCOVE (Kruschke, 1992).

A common method of representing stimuli is to treat each exemplar as a point in a multidimensional feature space. While there are an infinite number of potential visual and non-visual features that could be used, we assume that we can identify the subset of features that are most salient and relevant for the categorization task. The stimulus representation is denoted by $\mathbf{Z}$, where $\mathbf{z}_i$ indicates the D-dimensional feature vector of the ith stimulus. The input layer activations encode the query stimulus, $\mathbf{x} = \mathbf{z}_{query}$.

The similarity kernel $s(\mathbf{z}, \mathbf{z}')$ specifies how similarity between two stimuli ($\mathbf{z}$ and $\mathbf{z}'$) decays as a function of distance in feature space. The form of the similarity kernel is constrained by existing psychological theory. Following Roads and Mozer (2017), we integrate various psychological models (Jones, Love, & Maddox, 2006; Jones, Maddox, & Love, 2006; Nosofsky, 1986; Shepard, 1987) into a general form to obtain:

$$s(\mathbf{z}, \mathbf{z}') = \exp\left(-\beta \left\|\mathbf{z} - \mathbf{z}'\right\|_{\rho,\mathbf{w}}^{\tau}\right) + \gamma, \tag{1}$$

where β, ρ, τ, and γ control the gradient of generalization. The norm $\|\mathbf{z} - \mathbf{z}'\|_{\rho,\mathbf{w}}$ denotes the weighted Minkowski distance:

$$\|\mathbf{z} - \mathbf{z}'\|_{\rho,\mathbf{w}} = \left(\sum_{j=1}^{D} w_j \left| z_j - z'_j \right|^{\rho} \right)^{\frac{1}{\rho}} \text{ where } w_j \geq 0 \text{ and } \sum_{j=1}^{D} w_j = D.$$

The parameter ρ controls the type of distance (e.g., $\rho = 2$ yields Euclidean distance) and D indicates the dimensionality of the embedding. The weights w correspond to attention weights and allow the similarity kernel to model differences in how individuals or groups attend to different dimensions in the psychological embedding. The weights sum to D so that when all the weights are equal, i.e., $w_j = 1$, we recover the standard (unweighted) Minkowski distance. While the remainder of this work assumes this similarity kernel, other differentiable similarity kernels could be substituted without loss of generality. For convenience, the parameters ρ, β, τ, and γ are denoted by the set variable $\boldsymbol{\theta}$.

The activation of the ith hidden unit $h_{,i}$ is determined by the similarity kernel (Equation 1),

$$h_i = s\left(\mathbf{x}, \mathbf{z}_i\right). \tag{2}$$

The vector $\mathbf{z}_i$ specifies the location of a particular basis function. An RBFN network typically has multiple basis functions. The basis function locations can be determined in a number of ways, which are discussed shortly.

The hidden layer is connected to the output layer via a fully connected association weight matrix W. The output layer has the same number of units as the number of categories in the categorization task. A normalizing softmax operation is applied to the raw output activations to create output probabilities,

$$\mathbf{y} = \text{softmax}\left(\mathbf{h}W\right). \tag{3}$$

The EV of the query stimulus is the probability that the query is correctly categorized, i.e., the output probability of the query's category membership.

1.2 Joint inference of a psychological embedding and similarity kernel

In a typical category learning research paradigm, the stimulus representation Z is determined independently of the category learning model. For example, one could use a set of hand-coded features, low-level computer vision features, or features from pre-trained deep neural network to determine the stimulus representation. In this paradigm, the stimulus representation is then treated as fixed, and the remaining free parameters are fit using human training data.

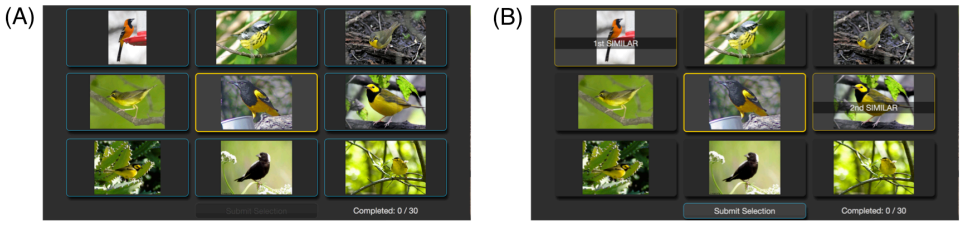

Figure 1: Sample displays shown to subjects. The center image is the query image while the surrounding images are the reference images. (**A**) Initially no reference examples are selected. (**B**) After participants make their selection, the two selected references are highlighted.

In contrast, the proposed approach determines both the stimulus representation and the parameters of the similarity kernel at the same time, without using human training data. This is achieved by applying an embedding algorithm to human similarity judgments in order to jointly infer the parameters of a similarity kernel and stimulus representation (Roads et al., 2018). While many algorithms exist for determining a stimulus representation, such as metric multi-dimensional scaling, non-metric multi-dimensional scaling, and t-distributed stochastic triplet embedding (Van Der Maaten & Weinberger, 2012), our approach leverages psychological theory to constrain the possible solutions. We refer to the inferred stimulus representation and similarity kernel as a *psychological embedding*. The procedure for collecting human similarity judgments and the mechanisms of the embedding algorithm are summarized below.

The first step to inferring a psychological embedding, is to collect human similarity judgments for the set of stimuli. Inspired by approaches used in the computer vision community (e.g., Wah et al., 2014), human similarity judgments are collected by having novice participants view displays composed of 9 randomly selected images arranged in a 3-by-3 grid (Figure 1). Each display is composed of a query image (center image) and eight reference images (surrounding images). Participants are asked to select the two reference images most similar to the query image. When participants make their selection, they also indicate which reference is most similar and second most similar. The ith judged display is denoted using a vector $\mathcal{D}_i = (q_i, a_i, b_i, c_i, d_i, e_i, f_i, g_i, h_i)$, where q_i is a scalar indicating the query image and a_i-h_i are scalars indicating the reference images. In this arrangement, a_i and b_i represent the most similar and second most similar references respectively. For convenience, $\mathcal{R}_i$ indicates the set of reference images of $\mathcal{D}_i$. The set of all judged displays is indicated by $\mathcal{D}$.

Next, the set of all judged displays $\mathcal{D}$ is used to jointly infer a stimulus representation and similarity kernel. Given a set of observations, the likelihood of the

data given the model parameters is:

$$\mathcal{L} = \prod_i p\left(\mathcal{D}_i | \mathbf{Z}, \boldsymbol{\theta}\right). \tag{4}$$

In the case where participants select and rank two reference images, the likelihood of a single judged display is:

$$p\left(\mathcal{D}_i^{(8r2)} | \mathbf{Z}, \boldsymbol{\theta}\right) = p\left(a_i | \mathbf{Z}, \boldsymbol{\theta}\right) p\left(b_i | a_i, \mathbf{Z}, \boldsymbol{\theta}\right). \tag{5}$$

The superscript $8r2$ on $\mathcal{D}_i$ indicates that participants select and rank two images from eight possible reference images.

Given a similarity kernel, the likelihood of subject selections are modeled in the same spirit as Luce's ratio of strengths formulation (Luce, 1959) where the probability of selecting a given reference is proportional to the similarity between the query and that reference:

$$p\left(\mathcal{D}_i^{(8r2)} | \mathbf{Z}, \boldsymbol{\theta}\right) = \frac{s\left(\mathbf{z}_{q_i}, \mathbf{z}_{a_i}\right)}{\sum_{r \in \mathcal{R}_i} s\left(\mathbf{z}_{q_i}, \mathbf{z}_r\right)} \frac{s\left(\mathbf{z}_{q_i}, \mathbf{z}_{b_i}\right)}{\sum_{r \in \mathcal{R}_{i \neg a_i}} s\left(\mathbf{z}_{q_i}, \mathbf{z}_r\right)}. \tag{6}$$

where $s\left(\mathbf{z}_i, \mathbf{z}_j\right)$ is the similarity kernel defined in Equation 1. During inference of the psychological embedding we assume a single set of attention weights and therefore set all attention weights to one.

In the case where participants only select one image from eight possible reference images, the likelihood of a single display is given by:

$$p\left(\mathcal{D}_i^{(8r1)} | \mathbf{Z}, \boldsymbol{\theta}\right) = \frac{s\left(\mathbf{z}_{q_i}, \mathbf{z}_{a_i}\right)}{\sum_{r \in \mathcal{R}_i} s\left(\mathbf{z}_{q_i}, \mathbf{z}_r\right)}. \tag{7}$$

If the observed data includes data from multiple display configurations, the likelihood for each display configuration can be multiplied together to give a combined likelihood. For example, if some displays had participants select and rank two references from eight choices while other displays had participants select one reference from eight choices the combined likelihood is:

$$\mathcal{L} = \prod_i p\left(\mathcal{D}_i^{(8c1)} | \mathbf{Z}, \boldsymbol{\theta}\right) \prod_j p\left(\mathcal{D}_j^{(8c2)} | \mathbf{Z}, \boldsymbol{\theta}\right). \tag{8}$$

After maximizing the log-likelihood using gradient decent, we obtain a stimulus representation and a corresponding similarity kernel that models human-perceived similarity.

1.3 Candidate RBFN implementations

After a psychological embedding has been inferred, an RBFN model is missing three pieces: the attention weights, the basis function locations and the association weight matrix. Existing *exemplar* and *prototype* models provide guidance on setting these parameters (e.g., Kruschke, 1992; Minda & Smith, 2002; Nosofsky, 1986; Smith & Minda, 1998). Each variant has different implications for how the basis function locations and association weight matrix are determined.

An exemplar model takes a fine-grained approach and places a basis function at the embedding location of each exemplar. The association weight matrix is fixed by assuming that each hidden until is only connected to the output unit that corresponds to its category. This version closely resembles Generalized Context Model (Nosofsky, 1986) except that there is no softmax free parameter that adjusts the determinism of human responses.

A prototype model takes a coarser approach and locates a basis function at the centroid of all exemplars belonging to a given category. In other words, a single basis function is used to represent each category. In a prototype model, we assume that the association weight matrix is the identity matrix that connects each category basis function to its appropriate output unit. The prototype implementation requires a bit more care to set up properly. Since an embedding is inferred on individual stimuli and the prototype basis function represents a category average, the parameters of the similarity kernel must be appropriately constrained and adjusted. First, the embedding algorithm is constrained to infer solutions where $\rho = 2$, $\tau = 2$, and $\gamma = 0$. Second, one multivariate Gaussian is fit for each category. The fitted Gaussians are then used as the corresponding basis function for each category. Since the fitted Gaussians are not constrained to be spherical, the basis functions are not radial basis functions. However, the additional flexibility was allowed to give the prototype model the best chance at making good predictions.

Following the approach of Nosofsky (1986), the attention weights w are set such that the RBFN model minimizes categorization error across all possible query stimuli. The reader may be wondering why the attention weights that were fixed during inference are now being altered. The embedding algorithm is blind to category membership and infers a category-agnostic representation. It is assumed that an expert would adjust their attention weights to maximize categorization performance given a set of category labels.

Before computing EVs, we must decide how much knowledge a given model posseses. If we use all of the stimuli to create basis functions, an exemplar model would predict that all stimuli are easy to categorize since the distance between itself and its corresponding basis function is zero. To prevent this uninteresting behavior, we employ a leave-one-out approach to compute EVs. For convenience,

let us denote the set of all stimuli less the ith stimulus as $\mathcal{I}_{\neg i}$. To compute the EV of ith stimulus, we instantiate an RBFN using the set $\mathcal{I}_{\neg i}$ to determine the locations of the basis functions. This design amounts to assuming a categorization model that has complete knowledge of all stimuli except the ith stimulus. Alternatively, a subset of the stimuli could be used, which would correspond to a model with only partial knowledge of the domain. However, we are primarily interested in computing EVs that are most likely to mimic ratings given by experts.

1.4 An alternative count based model

Instead of using a model constrained psychological theory, it is possible to employ a model that removes theory all together. In this approach, instead of inferring a psychological embedding from human similarity judgments, the similarity judgments are used to directly compute EVs. Each judged display tells us how often a given exemplar was judged to be similar to an exemplar of the same category and an exemplar of a different category. For example, if a participant participant sees a query stimulus belonging to category j and selects two references that also belong to category j, this provides two votes that the query stimulus is easy. If a participant sees another query stimulus belonging to category j, but selects two references that belong to category j and k respectively, this provides one vote that the query stimulus is easy. By looping over all judged displays, a simple count matrix can be assembled that tracks how often a given exemplar is judged to be more similar to a reference of the same category versus a reference of a different category. After looping through all judged displays, the count matrix can be normalized such that each row sums to one. Each row in the normalized count matrix gives the probability that a particular exemplar will be judged to more similar to a reference of the same category versus a reference of a different category. These probabilities can be used to estimate EV. Although simple, this *count-based* model serves as an important control that highlights the value, if any, of using a model that leverages an inferred psychological embedding.

2 Experiments

The above models encompass three intuitive methods for predicting difficulty. A good test of the proposed approach is to compare the predicted EVs to empirically derived EVs. Using two different human training experiments, we compute empirical EVs and compare them to the EVs predicted by three different models: an exemplar-, prototype-, and count-based model. The comparison will determine if the general approach is valuable and which model makes the most accurate predictions. For each experiment, we provide a brief description of the experimental

design followed by a comparison of the predicted and observed EVs.

2.1 Validation Experiment 1

The goal of the first validation experiment is to predict the EVs associated with categorizing a set of 156 bird images representing 12 different categories. Empirical EVs are derived from Experiment 1 of a previously conducted study (Roads & Mozer, in submissiona). The exemplar-, prototype-, and count-based models use similarity judgments that were previously collected for a similarity judgment database (Roads & Mozer, submitted). An abbreviated description of the human training study is included, with an emphasis given to details pertinent to the current work.

2.1.1 Methods

Participants Two sets of participants were used in this experiment. A psychological embedding was constructed from similarity judgments collected from 232 participants. Human training data was collected from 160 participants. All participants were recruited from Amazon Mechanical Turk (AMT) and paid at at rate of approximately $8.00 per hour for their time and effort.

Materials A set of 156 bird images, representing 12 different species (Figure 2A-C), were selected from the CUB-200 dataset (Wah, Branson, Welinder, Perona, & Belongie, 2011). For each species, 13 exemplars were hand picked by the first author. Exemplars were selected to make sure that the resolution of the image was sufficiently high, the bird was clearly visible in the image, and the bird exhibited the visual features characteristic of the species. To ensure a sufficiently challenging and representative task, species were selected such that there were three groups of four visually similar species.

Procedure Empirical EVs and predicted EVs were derived using two distinct procedures. The procedure for deriving empirical EVs from training data is described first. The procedure for computing EVs from the candidate models is described second.

During the training experiment, participants completed trials at their own pace for approximately one hour. At the beginning of the experiment the set of stimuli was randomly partitioned into a practice and assessment set. The practice set was composed of seven exemplars from each category, while the assessment set was composed of the remaining six exemplars from each category. The practice and assessment set were further partitioned into mini-sets of 12 exemplars containing

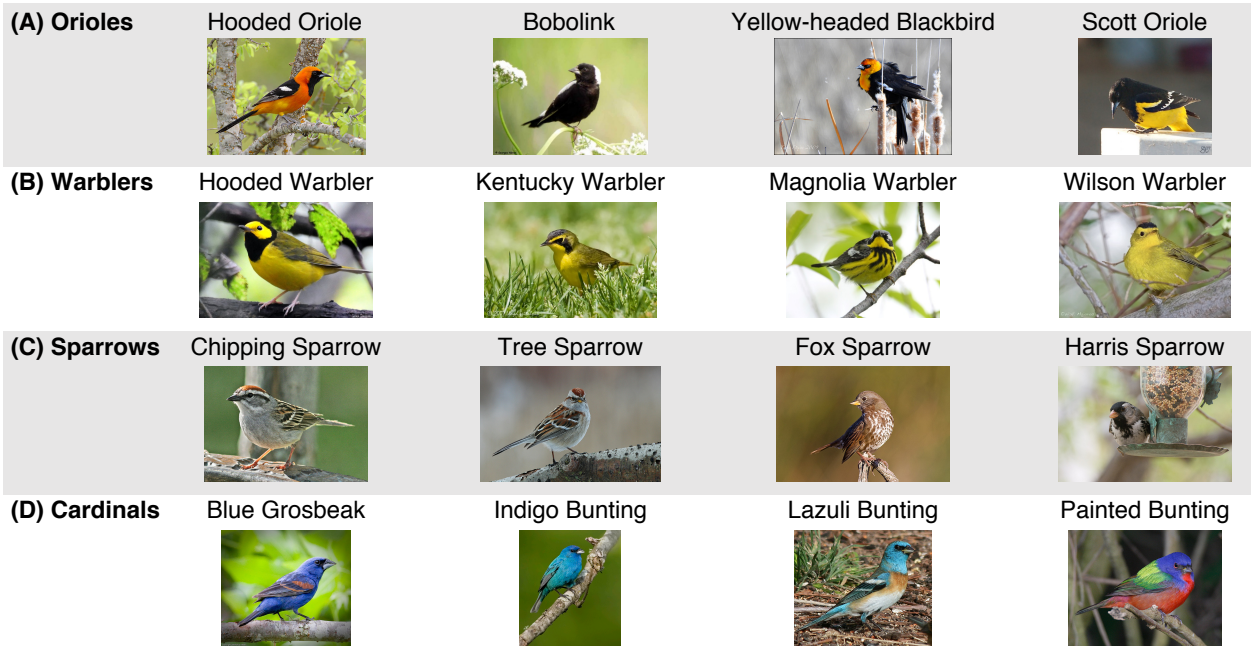

Figure 2: Example stimuli of the different bird species used in this work. Each row contains four similar bird species, each of which belong to the same or similar taxonomic family. Validation Experiment 1 used the 12 bird species in rows A-C. Validation Experiment 2 used all 16 bird species.

one exemplar from each category. Each mini-set was used to create a mini-block composed of 12 trials.

At the highest level, the experiment was composed of three parts. Each part consisted of a practice phase followed by an assessment phase. Each practice phase took a fixed time of 15 minutes. During a practice phase, trials were arranged into mini-blocks consisting of exemplars from the practice set. If a participant made it through all practice mini-blocks, the sequence of practice mini-blocks was repeated. Each assessment phase was composed of a fixed number of trials. During each assessment phase, two mini-blocks (24 trials) were shown to participants. Once the exemplars were shown during the assessment phase, they were added to the practice set for use in the next practice phase. Critically, all trials presented during the assessment phase used unseen stimuli. On all trials (both practice and assessment), a query stimulus was presented along with a blank response box. Participants typed the name of the category corresponding to the query stimulus. After submitting their response, participants received corrective feedback. Participants were not scored based on capitalization and answers within an edit distance of two were marked as correct.

In addition to this basic setup, some participants were assigned to conditions

where they could request clues on some of the practice trials. The clue-enabled trials are referred to as enriched trials. In the *standard* condition, participants were never given enriched training trials. In the *enriched-3* condition, participants were given enriched training trials in all three parts of the experiment. In the *enriched-2* condition, participants were given enriched training trials in the first two parts of the experiment. In the *enriched-1* condition, participants were given enriched training trials only on first part of the experiment. For the purpose of this work, we make no distinction between conditions and use all conditions in order to predict EVs. Each condition contained 40 participants.

Empirical EVs were computed using participant responses from the assessment trials only. For each query exemplar in the assessment trials, the exemplar ease value (EEV) was determined by counting the total number of times the query exemplar was categorized correctly and dividing it by the total number of times the it was shown. Each category ease value (CEV) was computed in an analogous way–by counting the number of times a category was classified correctly and dividing it by the number of times the category was shown.

Predicted EVs were obtained by inferring a psychological embedding from 7,520 2-choose-1 displays and 4,733 8-choose-2 displays. Two separate embeddings were inferred: one for the exemplar model and a second for the prototype model. The psychological embedding inferred for the exemplar model was subject to no constraints. The psychological embedding used for the prototype model was constrained to have a Gaussian similarity kernel in L_2 space in order to cohere with the later Gaussian fitting procedure used by the prototype-based RBFN. For visualization purposes, a two-dimensional embedding of the stimuli is provided in Figure 3A. After inferring the respective embeddings, the exemplar and prototype models were used to generate predicted EEVs. The raw similarity judgments were used by the count-based model to generate EEVs. Predicted CEVs were computed by taking the average EEV of all exemplars within a particular category.

Each model was evaluated by computing the Spearman rank correlation coefficient between the models predicted EVs and the empirical EVs. Separate correlations where computed for EEVs and CEVs. In order to determine the best model, the method of Steiger (1980) was used to test if the difference between the different correlation scores was significant.

2.1.2 Results

The primary hypothesis of this work is that simple models can be used to predict EVs for a set of stimuli. Three models were proposed as candidates, with the hypothesis that they would perform differently. Results indicated that simple models can be used to predict difficulty (Figure 4A). However, results were inconclusive regarding the best model. Tests of the three pair-wise comparisons

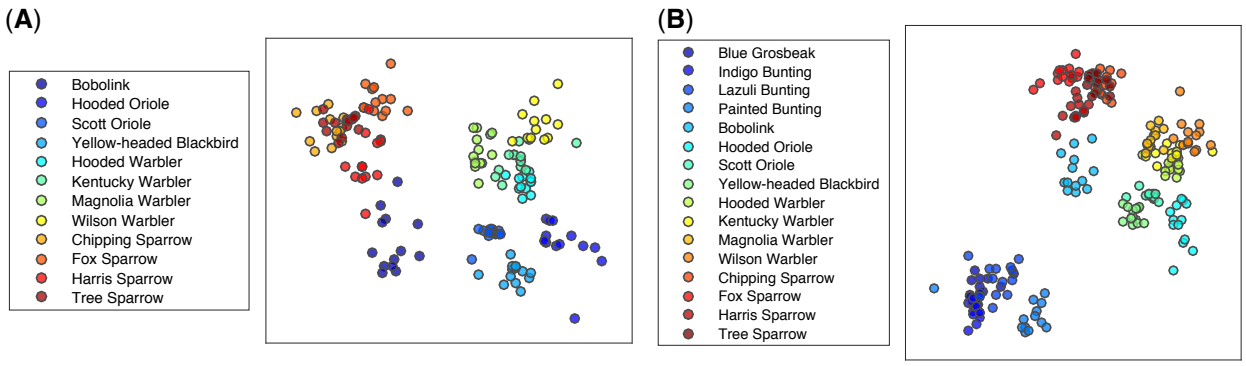

Figure 3: A two-dimensional psychological embedding of (**A**) the 156 unique bird images used in Validation Experiment 1 and (**B**) the 208 bird images used in Validation Experiment 2. Each point represents a unique stimulus and is colored according to category (i.e., species). Images judged to be similar are located closer together than images judged to be dissimilar.

were conducted using Bonferroni adjusted alpha levels of .0167 per test (.05/3). Results indicated that the correlation was significantly lower for the prototype model ($\rho_s = .42$) than both the exemplar ($\rho_s = .67$), $T_2(155) = 4.16$, $p < .001$ and count model ($\rho_s = .63$), $T_2(155) = 2.97$, $p = .003$. The pairwise comparison between the correlations for the exemplar and count model was non-significant. The same testing procedure was applied to the correlations between empirical category-level difficulty and predicted category-level difficulty. Results indicated that all pairwise comparisons were non-significant.

2.1.3 Discussion

The results of Validation Experiment 1 are encouraging because they suggest that simple models can be used to predict EVs. This is an exciting possibility given its potential usefulness in category training. However, the experiment did not clearly distinguish between the candidate models. Part of the problem concerns the data obtained from the human training experiment. The data collected from the human training experiment comes from assessment trials that occur near the beginning, middle, and end of the experiment. The EVs computed from the models assume a near-complete knowledge state. In principle, assessment trials from the last phase of the experiment could be used to derive empirical EVs. However, doing so significantly reduces the number of observations assocaited with each exemplar. When only a few observations are available, the EV estimate is very noisy. Instead, a new experiment is needed where all of the data is taken from assessment trials that occur at the end of the training period and where the number of observations

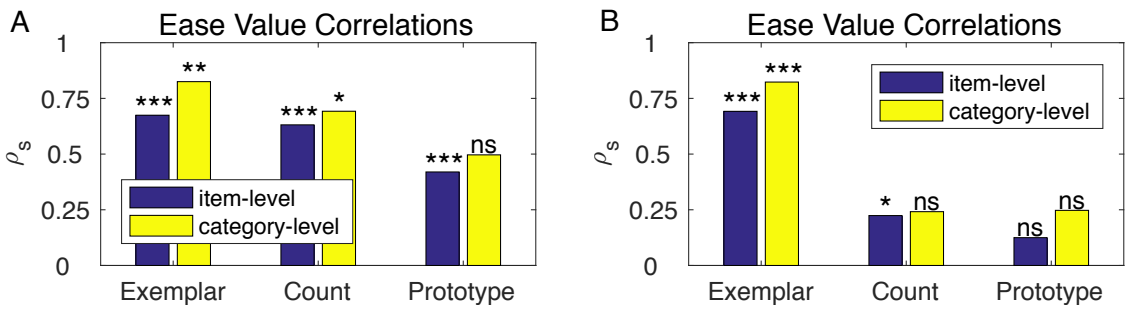

Figure 4: The Spearman's rank correlation coefficient between empirically derived EVs and predicted EVs. Correlations were computed for exemplar ease values (EEVs) and category ease values (CEVs). Symbols above the bars indicate the significance of the correlation. (**A**) Correlation coefficients using a dataset of 12 bird species composed of 156 unique images and behavioral data from 160 participants. (**B**) Correlation coefficients using a dataset of 16 bird species composed of 208 unique images and behavioral data from 120 participants.

is sufficiently large.

2.2 Validation Experiment 2

In the second validation experiment, the goal is to predict the EVs associated with categorizing a set of 208 bird images representing 16 different categories. Empirical EVs are derived from human training data collected as part of a previously conducted study (Roads & Mozer, in submissionb). The exemplar-, prototype-, and count-based models use similarity judgments that were previously collected for a similarity judgment database. An abbreviated description of the human training study is included, with an emphasis given to details pertinent to the current work.

2.2.1 Methods

Participants Two sets of participants were used for this experiment. A psychological embedding was constructed from similarity judgments collected from 342 participants. Human training data was collected from 120 participants. All participants were recruited from Amazon Mechanical Turk (AMT) and paid at at rate of approximately $8.00 per hour for their time and effort.

Materials A small dataset of 208 bird images representing 16 species was collected from the CUB 200 image dataset (Wah et al., 2011). Species were selected such that there were four groups of birds composed of four similar looking species,

roughly corresponding to four taxonomic families. For each species, 13 exemplars were hand picked by the lead author in same manner as described in Validation Experiment 1. The image dataset was an expanded version of the image dataset used in Validation Experiment 1 with four new species making up a new taxonomic family (Figure 2D).

Procedure Like Experiment 1, two distinct procedures were used to derive empirical EVs and a predicted EVs. The procedure used to derive the empirical EVs is described first. The procedure for computing EVs using the candidate models is described second.

The training experiment was split into a single *practice phase* immediately followed by an *assessment phase*. During practice phase, participants completed 224 practice trials in approximately 40 minutes. During the assessment phase, participants completed 96 trials in approximately 15 minutes. Including instructions and breaks, the entire experiment took approximately one hour.

The entire stimulus set was partitioned into a practice and assessment set. The practice set included seven randomly selected exemplars from each of the 16 categories. The assessment set included the remaining six exemplars from each category. The same partition was used for all participants.

During the practice phase, each exemplar in the practice set was shown twice. The practice phase was divided into four equal length blocks in order to allow participants to rest if desired. The order of the practice trials was determined by the experiment condition. During the assessment phase, participants saw each exemplar from the assessment set once. The assessment trials were organized into 6 mini-blocks such that each mini-block showed one exemplar from each category. Every participant saw the same randomly generated assessment sequence.

Every trial displayed a query stimulus and a text box for typing a response. On practice trials, a submitted answer was graded and corrective feedback was shown to participants. Participants chose when to advance to the next trial by clicking a button. Corrective feedback displayed the participant's response as well as the correct response. On assessment trials, no feedback was provided and participants clicked a button to advance to the next trial.

Participants were randomly assigned to one of three scheduling conditions that used a condition-specific scheduling policy for sequencing practice trials. For the purpose of this work, no distinction is made between the different conditions.

Empirical EVs were from participant responses on all assessment trials. For each query exemplar in the assessment trials, the EV was determined by computing the total number of times the query exemplar was categorized correctly and dividing it by the total number of times the query exemplar was shown. Since every participant used the same partition, empirical EVs could only be computed for 96 of the exemplars.

Predicted EVs were obtained and each model was evaluated using the same method from Validation Experiment 1. In this experiment embeddings were inferred from 7,520 2-choose-1 displays and 8,772 8-choose-2 displays. A two-dimensional embedding of the stimuli is provided in Figure 3B for visualization purposes. As before, the method of Steiger (1980) was used to test if the difference between dependent correlation scores was significant.

2.2.2 Results

Having already confirmed that simple models can be used to predict EVs, the primary hypothesis of Experiment 2 is that the models would perform differently. The primary hypothesis was confirmed and results indicated that the exemplar model was best at predicting EEVs and CEVs (Figure 4B).

Tests of the three pair-wise comparisons were conducted using Bonferroni adjusted alpha levels of .0167 per test (.05/3). Results indicated that the correlation was significantly higher for the exemplar model ($\rho_s = .69$) than both the count ($\rho_s = .22$), $T_2(95) = 5.24$, $p < .001$ and prototype model ($\rho_s = .12$), $T_2(95) = 6.19$, $p < .001$. The pairwise comparison between the correlations for the prototype and count model was non-significant.

The same testing procedure was applied to the correlations between empirical category-level difficulty and predicted category-level difficulty. Results indicated that the correlation was significantly higher for the exemplar model ($\rho_s = .82$) than both the count ($\rho_s = .24$), $T_2(15) = 3.44$, $p = .004$ and prototype model ($\rho_s = .25$), $T_2(15) = 4.62$, $p < .001$. The pairwise comparison between the correlations for the prototype and count model was non-significant.

2.2.3 Discussion

The results of the current experiment replicated the finding of Validation Experiment 1 that simple models can be used to predict EVs at both the exemplar and category level. Critically, the results of the current experiment convincingly demonstrate that the exemplar model is best able to predict EVs.

It is notable that the prototype model does so poorly in this experiment while it performed reasonably well in Experiment 1. One potential source of this discrepancy is that the empirical observations of Experiment 1 were derived from data that included relatively early assessment trials. It is possible that prototype models describe early, novice-like knowledge representations while exemplar models do a better job of capturing the minutia of expert knowledge representations (Smith & Minda, 1998). This perspective is consistent with computational models such as SUSTAIN (Love et al., 2004) and nonparametric Bayesian models of category learning (Sanborn, Griffiths, & Navarro, n.d.), which grow the complexity of the knowledge state as needed.

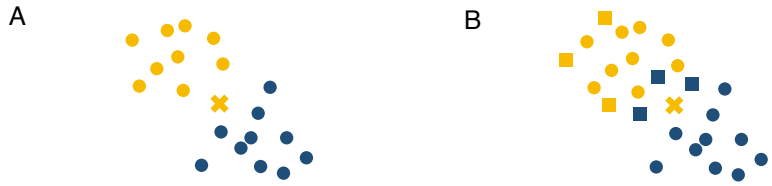

Figure 5: A cartoon depicting two different objective sets in a two-dimensional feature space. Each exemplar is represented by an icon where the color of the icon indicates the exemplar's category membership. (**A**) In the original objective set, the categories are non-overlapping. (**B**) After adding six new exemplars (indicated by squares), the categories are now slightly overlapping. The exemplar denoted by an "x" will be easier to categorize in the first objective set compared to the second objective set.

3 General Discussion

The proposed approach hinges on the fact that the free parameters of a simple category learning model can be fit using less expensive human data. Using a simple category learning model is made possible because we do not need to model trial-to-trial changes in knowledge, only the end state. This eliminates the need to include free parameters that govern state updating behavior.

In addition to the points already covered, the proposed approach is advantageous because it allows for extendabilty. In part, an EV is challenging to obtain because it is defined relative to a learning objective. An EV will depend on the stimulus itself as well as all the other stimuli defined to be part of the *objective set*, i.e., the set of all stimuli a learner is expected to know. To illustrate this point, consider a binary categorization task where each exemplar is characterized by two feature dimensions (Figure 5). In the first scenario, the two categories are non-overlapping (Figure 5A). In a second scenario, the objective set has been expanded by six exemplars to create overlapping categories (Figure 5B). The exemplar denoted by the gold "x" will likely be easier to categorize in the first scenario compared to the second scenario. This simple example illustrates how the EV of an exemplar fundamentally depends on the objective set and the category membership of its (potentially confusing) neighbors. The issue is further complicated by the fact that researchers rarely have access to the complete objective set. In theory, an objective set is composed of an infinite number of stimuli, with some stimuli only slightly different from others. In practice, researchers use an *experimental set*, a subset of the objective set that is assumed to mirror the EVs of the objective set. In an ideal scenario, the experimental set sufficiently captures the structure present in the objective set.

If a researcher decides to expand (or shrink) their stimulus set, the correspond-

ing EVs may change. If EVs were estimated empirically from error statistics, the corresponding behavioral data may not be useful anymore. In contrast, similarity judgments are robust to changes in the stimulus set. Even if the stimulus set changes, past similarity judgments are still usable. Since researchers are often faced with immense uncertainty, it is reassuring to know that similarity judgments will hold their value even if research goals and experimental designs change.

4 Conclusion

The goal of this work was two-fold. First, this work aimed to specify a formal and scalable approach for computing ease values at both the exemplar-level (EEV) and category-level (CEV). Scalability is achieved by fitting critical model parameters with human data collected from a cost-effective visual similarity task. Second, this work strived to provide convincing evidence that the proposed approach works well in practice. The results of two separate validation experiments show that an exemplar-based model is able to predict empirical EVs with high accuracy at both the exemplar-level and category-level. Importantly, the proposed approach can easily be used with real-world professional datasets, which we hope will enable the development of better skill assessment protocols and more efficient training systems.

Acknowledgments

This research was supported by NSF grants SES-1461535, SBE-0542013, SMA-1041755, and DRL-1631428.

References

Evered, A., Walker, D., Watt, A. A., & Perham, N. (2014). Untutored discrimination training on paired cell images influences visual learning in cytopathology. *Cancer cytopathology*, *122*(3), 200-210.

Hornsby, A. N., & Love, B. C. (2014). Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition*, *3*(2), 72-76. Retrieved from http://www.sciencedirect.com/science/article/pii/S2211368114000321 doi: http://dx.doi.org/10.1016/j.jarmac.2014.04.009

Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *32*, 316–332.

Jones, M., Maddox, W. T., & Love, B. C. (2006). The role of similarity in generalization. In *Proceedings of the 28th annual meeting of the cognitive science society* (pp. 405–410).

Khan, F., Mutlu, B., & Zhu, J. (2011). How do humans teach: On curriculum learning and teaching dimension. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 24* (pp. 1449–1457). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/4466-how-do-humans-teach-on-curriculum-learning-and

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

Lindsey, R., Mozer, M. C., Huggins, W. J., & Pashler, H. (2013). Optimizing instructional policies. In C. B. et al. (Ed.), (p. 2778-2786). La Jolla, CA: Curran Associates, Inc.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological Review*, *111*(2), 309-332.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: Wiley.

McLaren, I. P. L., & Suret, M. B. (2000). Transfer along a continuum: Differentiation or association? In L. R. Gleitman & A. K. Joshi (Eds.), *roceedings of the twenty-second annual conference of the cognitive science society* (pp. 340–345). Cognitive Science Society.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 275292.

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

Roads, B. D., & Mozer, M. C. (2017). Improving human-machine cooperative classification via cognitive theories of similarity. *Cognitive Science: An Multidisciplinary Journal*, *41*, 1394-1411. doi: 10.1111/cogs.12400

Roads, B. D., & Mozer, M. C. (in submissiona). Using enriched training environments for visual category training.

Roads, B. D., & Mozer, M. C. (in submissionb). Visual category training using structure-sensitive scheduling.

Roads, B. D., & Mozer, M. C. (submitted). Obtaining psychological embeddings through joint kernel and metric learning.

Roads, B. D., Xu, B., Robinson, J. K., & Tanaka, J. W. (2018). The easy-to-hard training advantage with real-world medical images. *Cognitive Research: Principles and Implications*, *3*(38). doi: 10.1186/s41235-018-0131-6

Salmon, J. P., McMullen, P. A., & Filliter, J. H. (2010). Norms for two types of manipulability (graspability and functional usage), familiarity, and age of acquisition for 320 photographs of objects. *Behavior Research Methods*, *42*, 82–95.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (n.d.). Rational approximations to rational models : alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144-67.

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

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411-1436.

Spiering, B. J., & Ashby, F. G. (2008). Response processes in information–integration category learning. *Neurobiology of Learning and Memory*, *90*(2), 330–338.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245-251.

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The Caltech-UCSD Birds-200-2011 Dataset* (Tech. Rep. No. CNS-TR-2011-001). California Institute of Technology.

Wah, C., Horn, G. V., Branson, S., Maji, S., Perona, P., & Belongie, S. (2014, June). Similarity comparisons for interactive fine-grained categorization. In *Computer vision and pattern recognition (cvpr).* Columbus, OH.