UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Learning New Categories for Natural Objects

Permalink
https://escholarship.org/uc/item/89b6z6ns

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

ISSN
1069-7977

Authors
Zou, Wanling
Bhatia, Sudeep

Publication Date
2021

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Learning New Categories for Natural Objects

Wanling Zou (wanlingz@sas.upenn.edu)
Department of Psychology, University of Pennsylvania
Philadelphia, PA 19104 USA

Sudeep Bhatia (bhatiasu@sas.upenn.edu)
Department of Psychology, University of Pennsylvania
Philadelphia, PA 19104 USA

Abstract

People learn new categories on a daily basis, and the study of category learning is a major topic of research in cognitive science. However, most prior work has focused on how people learn categories over abstracted, artificial (and usually perceptual) representations. Little is known about how new categories are learnt for natural objects, for which people have extensive prior knowledge. We examine this question in three pre-registered studies involving the learning of new categories for everyday foods. Our models use word vectors derived from large-scale natural language data to proxy mental representations for foods, and apply classical models of categorization over these vectorized representations to predict participant categorization judgments. This approach achieves high predictive accuracy rates, and can be used to identify the real-world settings in which category learning is impaired. In doing so, it shows how existing theories of categorization can be used to predict and improve everyday cognition and behavior.

Keywords: categorization; learning; distributional semantics; word vectors; exemplar model

Introduction

Categorization is one of the fundamental functions of human cognition (Ashby & Maddox, 2005; Estes, 1994, Smith & Medin, 1981), and much research has focused on developing models of how people learn new categories over multi-dimensional representations. Most of this work has been limited to perceptual, and usually highly abstract and artificial, stimuli (e.g. geometric figures and line drawings). Although this design choice gives researchers considerable control over the experimental stimuli, it also limits the application of categorization models to naturalistic domains of real-world importance (see Nosofsky et al., 2017, 2018b, c for an extended critique).

Consider, for example, the task of learning whether or not a food item contains a newly discovered nutrient. People may be given some examples of foods with and without the nutrient, but must extrapolate from these examples to make predictions for hundreds of other common food items. Good predictions lead to better food choices, and success at category learning has direct implications for health outcomes. Similar types of learning are at also play in other health settings (e.g. learning which activities transmit a newly discovered virus), in social domains (e.g. learning which movies are liked or disliked by a new friend), and in various economic and financial settings (e.g. learning which stocks will be adversely affected by a new political or social development).

All of these tasks involve category learning processes that operate over multidimensional representations. Although existing models of categorization have been shown to be successful at modeling category learning over such representations, applying these models to naturalistic tasks has been difficult. The reason for this involves the complexity of the mental representations that people have for natural objects. Category learning models often assume that categories are learned based on similarities between objects. Such similarities are easy to measure when the stimuli involve simple geometric shapes. But it is much harder to specify the multidimensional representations that guide and constrain similarity (and in turn, category learning) for natural objects.

Previous attempts at constructing multidimensional representations for natural objects have applied techniques such as multidimensional scaling to matrices of similarity ratings data (Nosofsky, 1992; Nosofsky et al., 2018c). Although these methods yield reasonable representations for natural objects, they are costly to implement, and are hard to extend to objects for which similarity data are not available. More importantly, human knowledge about natural objects is far richer than what can be captured by a small number of dimensions derived from similarity ratings.

In a recent line of work, researchers have been using representations derived from convolutional neural networks to predict human categorization decisions of natural images (e.g., Battleday et al., 2017, 2019; Guest & Love 2017; Peterson et al., 2016; Sanders & Nosofsky, 2020). The key idea is that machine learning models can be applied to large scale digital data, such as images obtained from the internet, in order to specify the representational structures for items and objects at play in naturalistic categorization tasks. These representations can serve as inputs into established category learning models to model *human* category learning and predict categorization performance.

This is a promising solution to the problem of naturalistic category learning, and our goal is to apply a variant of this solution to settings with concepts that are communicated to participants verbally, and thus can be proxied by representations obtained from large scale natural language data. For this purpose, we use word vector models, which have had great successes in modeling human similarity judgments for words (Landauer & Dumais, 1997; Jones &

Mewhort, 2007; Mikolov et al., 2013; see also Griffiths et al., 2007). Such models uncover multidimensional representations for words using cooccurrence statistics in natural language. As words that usually occur together, or occur in similar contexts, are given similar vector representations, similarity in word vector space accurately predicts human assessments of similarity. For this reason, vector similarity also predicts semantic memory search, free association, semantic priming, and other memory phenomena that are influenced by similarities in representation (see e.g. reviews in Günther et al., 2019; Jones et al., 2015; Mandera et al., 2017). The ability of word vector models to proxy mental representations for natural objects also makes them useful for modeling everyday cognition and behavior. Recent work has shown that models based on word vectors are able to accurately predict social judgments (Caliskan et al., 2017; Bhatia, 2017); probability judgments and forecasts (Bhatia, 2017); risk perception (Bhatia, 2019a); multiattribute decisions (Bhatia, 2019b; Bhatia & Stewart, 2018) and health judgments (Gandhi et al., 2020) (see also Bhatia et al., 2019 and Richie et al., 2019 for reviews and discussion of this work).

Building on these insights, we combine word vectors for natural objects with a widely used psychological model of categorization – the Generalized Context Model (GCM, Medin & Schaffer, 1978; Nosofsky, 1986). GCM has already been shown to be a good model for studying category learning of natural images with high-multidimensional image vectors (e.g., Battleday et al., 2017, 2019; Nosofsky et al., 2018a, b, 2019), and can be similarly used to study category learning of natural objects with highly multidimensional word vectors. In three pre-registered studies involving binary categories of foods, we test the adequacy and applicability of this combined GCM-word-vector approach in predicting human performance for both experimenter-constructed and naturally occurring categories. We also use this approach to identify the types of tasks in which category learning is impaired. Our results shed light on the generalizability of established psychological models of categorization to everyday categorization decisions, and evaluate the applicability of word vectors for modeling naturalistic categorization and other high-level cognitive phenomena.

Modeling Approach

Word Vectors

We used the pre-trained Word2Vec semantic space model (Mikolov et al., 2013) to obtain vector representations for food items. This model was trained on a large dataset of Google News articles (roughly 100 billion words in size with a vocabulary of three million unique tokens) using the continuous bag-of-words (CBOW) method (which predicts words from their neighbors) and the skip-gram method (which predicts neighboring words of a given word). These two methods allow words that appear in similar contexts and share related meanings to be located in close proximity in the resulting Word2Vec semantic space.

Each of the three million words and phrases in the Word2Vec vocabulary is described using a 300-dimensional vector. This vocabulary includes a large number of food items, which we use in our studies. Specifically, for each food item, we used the Word2Vec vector representation corresponding to the lower case of the food word (e.g. *walnuts*). We took the plural form only for foods that are normally consumed in bulk (e.g. *walnuts* and *blueberries*). Everything else was in the singular form. It is also worth noting that in the Word2Vec model, singular and plural forms have very similar vectors. Foods with multiple words in their names had each word separated by an underscore (e.g. *cream_cheese*). Although the vectors in the original Word2Vec vocabulary have different magnitudes, we normalized all vectors to unit norm prior to analysis.

Generalized Context Model

The Generalized Context Model (GCM) assumes that categorization decisions are made based on the overall similarity between the to-be-classified object and all exemplars within the category. Although attention weights on dimensions are often included in GCM, we assumed equal attention weights on all 300 dimensions, since it is computationally intractable to fit attention weights on such high-dimensional data. We fit a simplified GCM with two free parameters – a sensitivity parameter, *c*, and a response-scaling parameter, γ. This model specifies the similarity between objects *i* and *j* as:

$$s_{ij} = e^{-cd_{ij}} \quad (1)$$

Additionally, in a binary categorization task, it specifies the probability of classifying object *i* in category A (vs. B) as:

$$P(A) = \frac{(\sum_{j \in A} s_{ij})^{\gamma}}{(\sum_{k \in A} s_{ik})^{\gamma} + (\sum_{k \in B} s_{ik})^{\gamma}} \quad (2)$$

Here the probability of categorizing an object *i* as A is equal to the summed similarity of *i* to all exemplars of category A divided by the summed similarity of *i* to all exemplars in both categories A and B (Equation 2). The similarity of objects *i* and *j* is an exponential decay function of the Euclidean distance between *i* and *j*, d_{ij} (Equation 1). In our case, we measure d_{ij} using the distance between unit normalized word vectors on the 300-dimensional Word2Vec space. The sensitivity parameter, *c*, determines how sensitive perceived similarity is with respect to change in the distance in the space. The response-scaling parameter, γ, measures the degree of determinism in the participant responses.

We evaluated the GCM model on a category learning task which offered participants a set of training examples, with objects categorized as A and B. This was followed by a testing phase in which participants categorized other objects as A or B without feedback. We estimated model parameters by minimizing the summed deviations between predicted categories and true categories in the training data. To find the best-fitting parameters of GCM, we used the Python function

Table 1: Three pairs of category structures generated by three different methods, food items that are closest to the category prototypes, and example foods in different categories.

Study	Generating Method	Category Structure	Prototypical food in "*With Nutrient X*"	Prototypical Food in "*Without Nutrient X*"	Example Food "*With Nutrient X*"	Example Food "*Without Nutrient X*"
1	K-means clustering	$k = 2$	Pear	Shrimp	Apricot	Chicken liver
		$k = 10$	Asparagus	Tomato	Coconut	Turkey
2	Shepard et al. (1961)	Type I	Broccoli	Tomato	Rabbit	Abalone
		Type V	Shiitake mushroom	Tomato	Peanut	Sour cream
3	Nutrient	Cholesterol	Shrimp	Mango	Snail	Tofu
		Lutein	Tomato	Shiitake mushroom	Eel	Trout

scipy.optimize.fmin which minimizes summed deviations using the downhill simplex algorithm with 100 iterations and random starting points for each iteration. We also fit a prototype analogue of GCM, in which classification probability was given by:

$$P(A) = \frac{{s_{iA}}^{\gamma}}{{s_{iA}}^{\gamma} + {s_{iB}}^{\gamma}} \quad (3)$$

Here s_{iA} is the similarity between object *i* and a prototype of category A. s_{iB} is likewise the similarity between object *i* and a prototype of category B. We calculated the vector representation for the prototype of a category by taking the arithmetic mean of the word vectors of all training exemplars in that category. After training the models on the training items, we applied the best-fitting parameters to predict classification probabilities and category labels for the test items. We evaluated our model fits by comparing the proportion of correct responses predicted by the models with the observed proportion of correct responses from human participants. We also calculated the Pearson correlation between item-level classification probabilities predicted by the models and the observed classification probabilities.

Before proceeding, it is useful to note that we did not fit our models to subject-level responses. The training task was passive, and model parameters were based only on the training data. It is likely that some of these parameters vary across participants, reflecting individual difference in noise, response scaling, and other variables.

Methods

Participants

We recruited a total of 302 participants – 101 participants (mean age = 32.26, 51.49% were female) in Study 1, 100 participants (mean age = 34.68, 44% were female) in Study 2, and 101 participants (mean age = 33.31, 41.58% were female) in Study 3 from Prolific Academic. All participants were from the U.S. and had an approval rate of 80% or above. They were paid at a rate of approximately $11 per hour.

Stimuli

The stimuli were 100 food items. We chose food to test out our approach because this is a domain that most people have extensive knowledge about. The Word2Vec model has also been shown to provide accurate representations for foods (e.g. Gandhi et al., 2020; Richie et al., 2019). To generate new binary categories, we used three different methods, each of which yielded one simple and one complex category structure. The simple and complex categories were offered to participants in between-participant conditions in the three studies. For all studies, we labeled the resulting binary categories as either *with nutrient X* or *without nutrient X*.

In Study 1, we applied k-means clustering with $k = 2$ and $k = 10$ on the Word2Vec vectors of the 100 food items. In the $k = 2$ case, we simply classified one of the clusters as corresponding to one of the categories, and the other cluster as corresponding to the other category. In the $k = 10$ case, we divided ten clusters into two equal-size categories in a way that maximized the average pairwise distance between cluster centroids that were grouped into the same category. This method yielded a complex category structure on the word vector space. In both cases, each of the resulting categories had 50 food items and category labels (*with nutrient X* or *without nutrient X*) were assigned at random.

In Study 2, we adopted two category structures from Shepard et al. (1961). Shepard et al., showed that all possible binary classifications of eight stimuli defined in a three-dimensional binary-value space can be summarized into six basic types (Figure 1). We decided to use the type I and type V category types from Shepard et al., as prior empirical work has shown that people consistently perform worse at type V than at type I. We constructed these two category structures by first reducing the 300-dimensional Word2Vec vectors to three dimensions using principle components analysis. Then we separated stimuli along each dimension by the median value of that dimension. This procedure resulted in eight regions of three-dimensional space (corresponding to the eight points described in Shepard et al. (1961)). Using the scheme illustrated in Figure 1, we generated the two binary category structures corresponding to the type I (simple) and type V (complex) problem. The category labels were then assigned randomly. Note that the resulting categories were

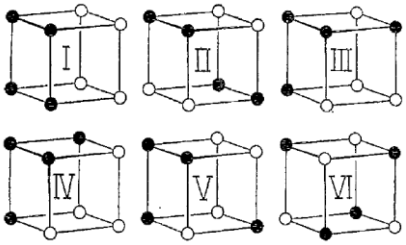

Figure 1: Schematic illustration of the six-type problems from Shepard et al. (1961). Each dot represents a stimulus and dots of the same color fall into the same category. Figure is from Shepard et al. (1961).

not balanced – the type I structure had 52 food items classified as *with nutrient X* and type V structure had 49 food items classified as *with nutrient X*. These minor differences should not influence our results.

In Study 3, we created two category structures based on whether foods have a certain real-world nutrient. Foods can be classified as either having cholesterol or not (a simple category structure, as most plant-based foods don't have cholesterol). They can also be classified as either having lutein or not (a complex category structure as there are some animal products that have lutein and some animal products that do not). There were 46 food items with cholesterol and 50 food items with lutein. They were labeled as *with nutrient X* in the simple and complex conditions respectively. The remaining foods were labeled as *without nutrient X*. Participants were not told that the nutrient was actually cholesterol or lutein. Table 1 summarizes the three generating methods, with the top item in each study corresponding to the simple structure and the bottom item in each study corresponding to the complex structure. Table 1 also provides examples of foods in different categories as well as foods whose word vectors are the closest to the word vectors of the category prototypes.

Procedure

Participants were randomly assigned to either the simple condition (k = 2 condition in Study 1, type I condition in Study 2, and cholesterol condition in Study 3), or the complex condition (k = 10 condition in Study 1, type V condition in Study 2, and lutein condition in Study 3). At the beginning of each condition in each study, we gave participants the following instruction: "*We recently discovered a nutrient X that may be found in some food items. In this study, we will show you food items that do or do not have this nutrient. Your task is to predict whether some other food items have this nutrient or not.*" Participants were then shown 50 training examples and asked to make predictions for 50 test examples. We also incentivized participants to make good predictions by paying an additional $1 to those whose predictive accuracy on the test items achieved the top 10% among all participants.

In each study, 50 foods (25 in each category) were training items. The remaining foods were test items. Training and tests items were the same for all participants in a given study. Thus, participants in both the simple and the complex conditions received the same training and test sets (though of course the category labels assigned to items in these sets changed based on the condition). This was done to ensure that item-specific effects in training and test sets did not influence our results.

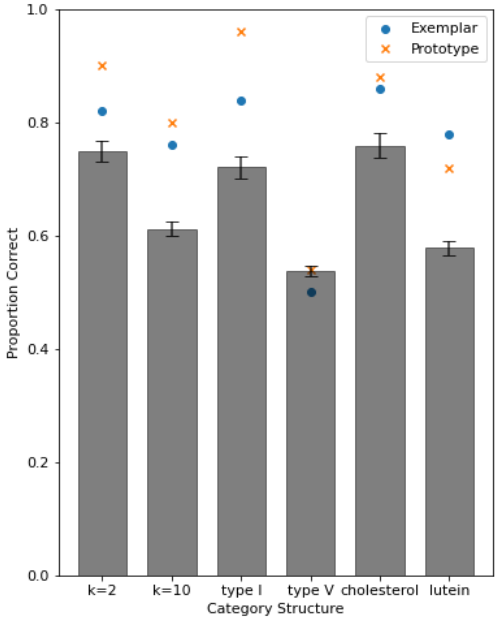

Figure 2: Mean observed proportions of correct responses in the three pairs of category structures as summarized in Table 1. Blue dots represent predicted proportions of correct responses by GCM (labeled "Exemplar"). Orange crosses represent predicted proportions of correct responses by the prototype analogue of GCM (labeled "Prototype"). Error bars represent standard errors.

We used an observational training procedure in which training items were shown on the same screen as the test items. Specifically, we presented participants with two tables, which listed the 25 training items in each of the two categories. We then asked participants to categorize the remaining 50 test items one at a time. The order of the 50 test items was randomized. The observational training procedure yields similar response patterns to a standard training procedure (in which participants respond to training stimuli one at a time following a corrective feedback until a certain number of correct responses are recorded successively) (Estes, 1994). Moreover, compared to the traditional paradigm, the observational paradigm is more similar to real-world environments where people have access to previously encountered exemplars when categorizing a new object. All three studies were pre-registered at aspredicted.org.

Results

Our three studies allow us to test whether or not GCM applied to word vector representations for foods is a good model of

how people learn new category labels for foods. Study 1 tests this in a setting in which the category labels involve either simple or complex boundaries on the underlying word vector space. If this space proxies people's actual mental representations for foods, then it should be easier for our participants to learn category labels in the $k = 2$ condition than the $k = 10$ condition. Study 2 attempts a similar test, but instead uses previous findings regarding the types of category structures that are easy or hard to learn by human subjects. Again, if the Word2Vec space proxies people's mental representations for foods, then it should be easier for our participants to learn category labels in the type 1 condition than the type V condition. Finally, Study 3 uses real-world categories that we expect to be easy or hard to learn. If the Word2Vec provides good representations for foods then we should observe higher accuracy rates for both our participants and our models in the cholesterol condition relative to the lutein condition.

In Figure 2, we plot mean observed and model-predicted proportions of correct responses on the test data in all three studies. Here we see that human participants typically performed better in the simple category condition ($k = 2$ condition in Study 1, type I condition in Study 2, and cholesterol condition in Study 3) than the corresponding complex condition ($k = 10$ condition in Study 1, type V condition in Study 2, and lutein condition in Study 3). We also conducted t-tests to evaluate these differences. These tests show that mean participant accuracy was significantly higher when the category structure was simple than when it was complex ($t(49,50) = 6.15$ in Study 1; $t(49,49) = 8.5$ in Study 2; and $t(48,51) = 7.41$ in Study 3, all $p < 10^{-5}$).

Importantly the GCM (exemplar) model mimicked these patterns, and achieved higher accuracy rates in the simple condition relative to the complex condition. The prototype analog of GCM also captured this trend in all studies. Note that both GCM and its prototype analogue fail to accurately predict precise proportions of correct responses by human participants. In most conditions, both models overpredict human accuracy, except in the type V condition where GCM underpredicts accuracy. This is due to the fact that we did note calibrate our model parameters on human data. Rather the parameters were fit to optimize performance on the training items, and made purely out-of-sample predictions for the test items. In this way, the models did not reflect participant tendencies. If we had, for example, fit the response scaling parameter to optimize fit to the participant data, the GCM model would likely have reflected a higher degree of noise and thus would have generated predictions closer to the observed participant accuracy. Nonetheless, the finding that the directional predictions are accurate across all three studies provides supporting evidence for our modeling framework.

Another way to test GCM (and its prototype analogue) is to compare the observed probabilities of a food item being classified in a given category against the predicted classification probabilities by the model. This is shown in Figure 3. Each point in each scatterplot represents a test item;

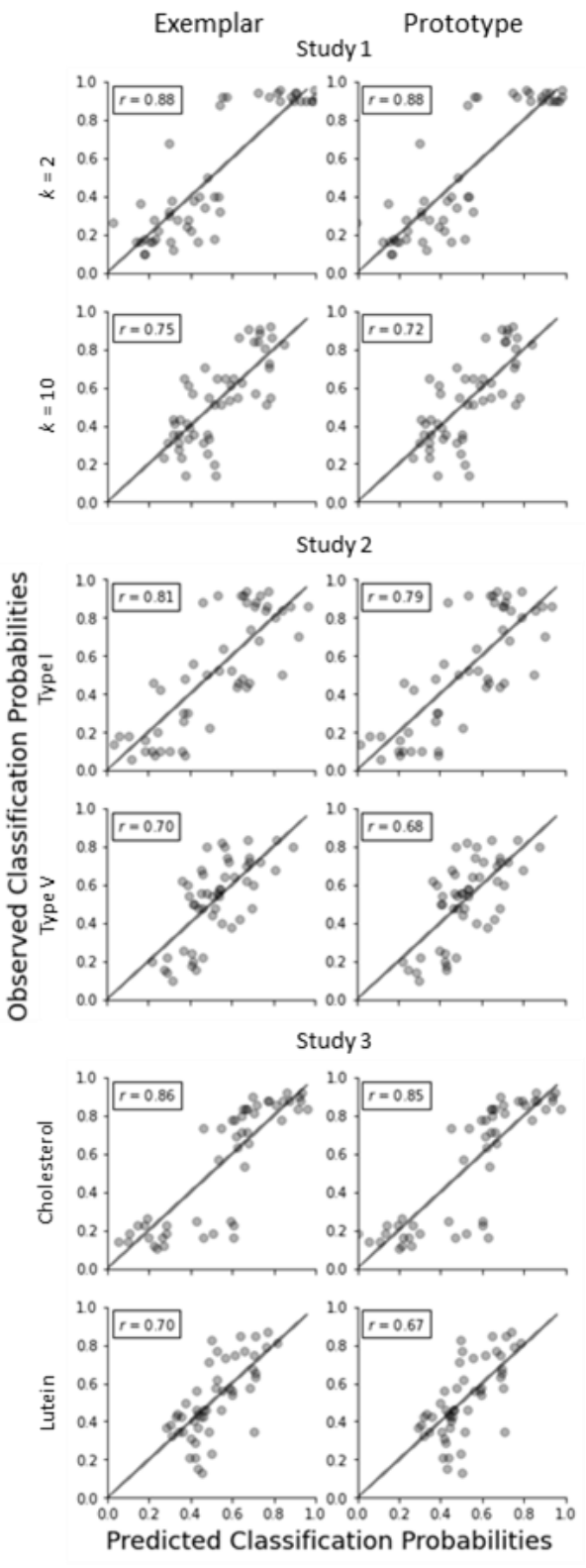

Figure 3: Scatterplots of observed probabilities of food being classified as "*without nutrient X*" vs. predicted classification probabilities by GCM (left column) and by the prototype analogue of GCM (right column) in the three pairs of category structures as summarized in Table 1, along with Pearson correlations.

thus, there are 50 points in each plot. For each category structure, we used separate OLS regressions to calibrate the classification probabilities from the outputs of Equation 2 (with the best-fitting parameters based on the training data) to human data. The left column shows the results of GCM (exemplar); the right column shows those of GCM's prototype analogue. Here we see high correlations for the GCM model (above $r = 0.70$), and somewhat lower correlations for the prototype analogue. The Pearson correlations are all significant (all $p < 10^{-6}$). Six paired t-tests show that the sum of squared residuals of GCM are significantly lower than those of its prototype analogue in the $k = 10$, type I and lutein condition ($p < 0.05$), suggesting GCM predicts the item-level classification probabilities much better than its prototype analogue in these conditions. When we control for the correct responses in multivariate regression analysis, the Pearson correlations are still all significant (all $p < 10^{-6}$), indicating that GCM and its prototype analogue predict human classification probabilities for both categories in consideration. However, the sum of squared residuals of GCM are significantly lower than those of its prototype analogue in the $k = 10$ and lutein condition ($p < 0.05$).

Discussion

We have proposed a modeling approach to study category learning for natural objects. Our approach uses word vector representations derived from large-scale natural language data as inputs into a psychological model of categorization. In three pre-registered studies, we evaluated this combined approach for both experimenter-generated and real-world categories of food items. We showed that the combined GCM and word vector model provided a good account of human data. Both GCM and its prototype analogue were able to capture observed patterns of accuracy in six category structures and accurately predicted aggregate human accuracy and item-level classification probabilities. The success of our approach suggests that word vector representations approximate human knowledge well and GCM provides a reasonable account for category learning. Although at such an early stage of the project, we have only considered a simplified version of GCM with two free parameters assuming equal attentional weights, future research should examine a more complex model that is trained on individual responses when corrective feedback is provided and holds different probabilistic assumptions for attentional weights. In addition, even though we did not intend to contrast exemplar and prototype models, our results appeared to support an exemplar-based classification process, as GCM consistently outperformed its prototype analogue. This finding replicates recent results of Nosofsky and colleagues (e.g. Nosofsky et al., 2018b, 2020) who find strong support for exemplar-based process in learning rock categories.

By integrating cognitive models of categorization and word-vector-based representations, our combined approach not only enriches our theoretical understanding of category learning, but also provides opportunities for real-world applications in many domains. For example, if we can predict how humans learn new categories of food items, we can identify food items that are easily misclassified or categories that are harder to learn, and subsequently improve health and risk communication by focusing on these food items or categories in public health campaigns. Similar applications are also possible for other domains in which new categories are learnt over (already known) natural objects. As noted by Goldstone (1994) everyday categories can range from natural kinds (e.g., animals, plants), to man-made objects (e.g., furniture, vehicles), to ad hoc categories (e.g., occupations that will likely be replaced by machines in the future), and to abstract concepts (e.g., food sources containing a particular nutrient). Thus, applying this combined approach to study everyday categorization will provide opportunities for real-world applications in many domains of policy and commercial relevance. Future research should extend this combined approach to more diverse category structures and other naturalistic domains. Furthermore, combining theories from cognitive psychology with new methods in machine learning and computational linguistics will open up a range of new research questions including how natural categories are represented in human mind, how knowledge of these categories is retrieved, and how new categories emerge.

References

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol.*, *56*, 149-178.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2017). Modeling human categorization of natural images using deep feature representations. *arXiv preprint arXiv:1711.04855*.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2019). Capturing human categorization of natural images at scale by combining deep networks and cognitive models. *arXiv preprint arXiv:1904.12690*.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20.

Bhatia, S. (2019a). Predicting risk perception: New insights from data science. *Management Science, 65*(8), 3800–3823.

Bhatia, S. (2019b). Semantic processes in preferential decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(4), 627–640.

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences, 29*, 31–36.

Bhatia, S., & Stewart, N. (2018). Naturalistic multiattribute choice. *Cognition*, *179*, 71–88.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186.

Estes, W. K. (1994). *Classification and cognition*. Oxford University Press.

Gandhi, N., Zou, W., Meyer, C., Bhatia, S., & Walasek, L. (2020). Knowledge representations in health judgments.

Proceedings of the 42nd Annual Conference of the Cognitive Science Society, 280–286.

Guest, O., & Love, B. C. (2017). What the success of brain imaging implies about the neural code. eLife, 6, e21397.

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition, 52*(2), 125–157.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211–244.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science, 14*, 1006–1033.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological review*, *114*(1), 1–37.

Jones, M. N., Willits, J., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of mathematical and computational psychology* (pp. 232–254). New York, NY: Oxford University Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207–238.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual review of Psychology*, *43*(1), 25–53.

Nosofsky, R. M., Meagher, B., & Kumar, P. (2020). Contrasting exemplar and prototype models in a natural-science category domain. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 641–647.

Nosofsky, R. M., Sanders, C. A., Gerdom, A., Douglas, B. J., & McDaniel, M. A. (2017). On learning natural-science categories that violate the family-resemblance principle. *Psychological Science, 28(1)*, 104–114.

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018a). A formal psychological model of classification applied to natural-science category learning. *Current Directions in Psychological Science, 27*, 129–135.

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018b). Tests of an exemplar-memory model of classification learning in a high-dimensional natural-science category domain. *Journal of Experimental Psychology: General, 147*, 328–353.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018c). Toward the development of a feature-space representation for a complex, natural-category domain. *Behavior Research Methods, 50*, 530–556.

Nosofsky, R. M., Sanders, C. A., Zhu, X., & McDaniel, M. A. (2019). Model-guided search for optimal training exemplars in a natural-science category domain: a work in progress. *Psychonomic Bulletin & Review, 26*, 48–76.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2016). Adapting deep network features to capture psychological representations. *arXiv preprint arXiv:1608.02164*.

Richie, R., Zou, W., & Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology*, *5*(1), 50.

Sanders, C. A., & Nosofsky, R. M. (2020). Training deep networks to construct a psychological feature space for a natural-object category domain. *Computational Brain & Behavior*, 1–23.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1–42.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts* (Vol. 9). Cambridge, MA: Harvard University Press.