Using Deep-Learning Representations of Complex Natural Stimuli as Input to Psychological Models of Classification

Craig A. Sanders (craasand@indiana.edu)
Robert M. Nosofsky (nosofsky@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 E. Tenth Street, Bloomington, IN., 47405 USA

Abstract

Tests of formal models of human categorization have traditionally been restricted to artificial categories because deriving psychological representations for large numbers of natural stimuli has been an intractable task. We show that deep learning may be used to solve this problem. We train an ensemble of convolutional neural networks (CNNs) to produce the multidimensional scaling (MDS) coordinates of images of rocks. We then show that not only are the CNNs able to predict the MDS coordinates of a held-out test set of rocks, but that the CNN-derived representations can be used in combination with a formal psychological model to predict human categorization behavior on a completely new set of rocks.

Keywords: deep learning; multidimensional scaling; categorization; psychological representations

Introduction

Numerous sophisticated formal models of human category learning and representation have been proposed in the field of cognitive science (for a comprehensive review, see Pothos and Wills, 2011). However, almost all rigorous quantitative tests of such models have been in highly simplified domains involving artificial category structures tested in laboratory experiments. In recent work, Nosofsky, Sanders and McDaniel (2018; see also Nosofsky, et al., 2017a) scaled up the application of such models by testing their ability to account for learning and generalization of rock classifications in the geologic sciences. Rock categories provide good examples of complex, high-dimensional category structures found in the natural world, so provide an intriguing and challenging test of the candidate models in the field.

Nosofsky et al.'s (2018) study focused on a well-known exemplar model of classification known as the generalized context model (GCM; Nosofsky, 1986). According to the GCM, people represent categories by storing individual exemplars of the categories in memory, and classify objects based on their similarity to the stored exemplars.

To apply the GCM, one needs to specify the multidimensional feature space in which the to-be-classified objects are embedded. In numerous past tests of the model, the derivation of the feature space was straightforward, because the objects used in the artificial category-learning experiments were simple stimuli composed of a small number of highly salient dimensions (e.g., geometric forms varying in shape, color, angle, and so forth). In a real-world category domain such as rocks, however, the derivation of the feature space becomes a highly ambitious task. The stimuli that compose such categories vary along a very large number of dimensions, many of which may be difficult to discern.

Thus, as a prerequisite to testing the exemplar model in the rock-classification domain, Nosofsky et al. (2017a) and Nosofsky, Sanders, Meagher and Douglas (2017b) engaged in extensive similarity-scaling studies of the rock stimuli. In these studies, observers provided similarity judgments among pairs of items drawn from a set composed of 360 rock pictures (10 categories of each of the broad divisions of igneous, metamorphic and sedimentary rocks, with 12 samples of each of the categories). Multidimensional scaling (MDS) (Shepard, 1980) was then used to model the similarity judgments to derive the rock feature space. In brief, in MDS, each object is represented as a point in a multidimensional space, with similarity presumed to be a decreasing function of distance in the space. A virtue of the MDS technique is that beyond summarizing large sets of similarity-judgment data, one can inspect the derived space to determine the psychological dimensions that compose the objects.

In the case of the MDS analysis of the rocks, the results were remarkably straightforward: An 8-dimensional solution provided a good account of the similarity structure of the 360 rock tokens that composed the 30 rock categories, and the derived dimensions could be interpreted in terms of: lightness/darkness of color, average grain size, shininess, roughness/smoothness, organization, chromaticity, hue, and shape-related components. Displays of the derived MDS solution are provided in the website (https://osf.io/w64fv/) associated with Nosofsky et al.'s (2017b) study. Perhaps most important, when used in combination with the MDS solution, the GCM was able to achieve good first-order quantitative predictions of rock-classification learning and generalization across a variety of conditions in which the nature of the training exemplars was manipulated (for details, see Nosofsky et al., 2018; for related work in the domain of semantic classification, see, e.g., Storms et al., 2000).

Despite its virtues, the MDS approach also has some limitations. One limitation is a practical one: In situations involving the scaling of large numbers of stimuli, deriving MDS solutions from similarity-judgment data requires the collection of a prohibitive amount of empirical data—for example, there are over 100,000 cells in the 360x360 similarity-judgment matrix used in Nosofsky et al.'s (2017b) study. If the goal is to position even larger numbers of stimuli in the high-dimensional feature space using these techniques, then the traditional approach becomes intractable.

Thus, in the present work our goal was to begin to test automated methods for deriving the natural-category feature space. Our key idea involves a novel integration in which MDS methods are combined with the use of deep learning convolutional neural networks (CNNs; e.g., Lecun et al., 2015). As is well known, CNNs have been used successfully to predict the classification of natural images from large data bases. In a typical CNN architecture, elementary visual inputs are converted to higher-order features via connections to a series of hidden convolutional layers and pooling layers, which then feed into fully connected layers and a final output layer that generates the classification responses. Recent research has shown that unlike classic computer vision algorithms, CNNs can be used to predict human category and typicality judgments regarding visual stimuli (e.g., Lake et al., 2015). Other work has advanced the idea that the deep features extracted after training the networks to predict visual categories can serve as candidates for the psychological feature-representations of the stimuli. Those deep-level features can then be used to predict human similarity judgments (Peterson et al., 2017; see also Rumelhart & Todd, 1993) or used as input to psychological models of classification (Battleday, Peterson, & Griffiths, 2017).

Despite these preliminary successes, the extent to which CNNs truly capture the detailed nature of human classification learning remains unknown. Thus, in the present work, we adopt an approach that is complementary to the past applications. Rather than training CNNs to classify objects into categories, we instead train them to predict the dimension values of individual exemplars derived from traditional MDS methods. Once the CNN is trained in this manner, new stimuli can be presented to the CNN and it can be used to automatically produce the coordinate values of the stimuli in the multidimensional psychological feature space. Thus, an unlimited number of stimuli from complex naturalistic domains can be scaled in this manner. The derived coordinate values can then be used in combination with formal models such as the GCM to predict categorization. In the remainder of this article, we explain the proposed procedure in depth, and present preliminary tests of the approach in the domain of rock classification.

Deep Learning Procedure

The basic plan of action for our deep learning procedure was to train CNNs to take images of rocks as input and yield their psychological representations as output. In this section we describe the specific data set, CNN architecture, and training procedure that we used. All procedures described in this section were implemented using the Keras Python package and Tensorflow (Abadi et al., 2016).

Data Set

We used Nosofsky et al.'s (2017b) data set to train our CNNs. To reiterate, this data set consists of 360 images of rocks belonging to 30 different categories along with each rock's 8-dimensionsal MDS coordinates. While the naïve approach would be to train and evaluate each network using all 360 images, CNNs may have millions of trainable parameters, and thus are prone to overfitting to noise and failing to generalize to new data. Therefore, we needed a means to compare the CNNs' generalization performance and not just their training performance. To this end, we split the data into three separate sets: a training set, a validation set, and a test set. CNNs were trained to minimize error on the training set, and each network's error on the validation set was computed to find the CNNs with the best generalization performance. Finally, these networks' error on the test set was computed to avoid overfitting to the validation set and to gain an unbiased estimate of their ability to generalize to completely new data. The training set was formed by randomly sampling 6 of the 12 rock tokens in each category, and the remaining tokens were evenly split between the validation and test sets. Therefore, there were 180 images in the training set, and 90 images in both the validation and test sets.

CNN Architecture

Our rocks data set is quite small for a deep-learning data set. By comparison, deep CNNs are often trained to perform image classification on the ILSVRC data set, which consists of over one-million images belonging to 1000 different categories (Russakovsky et al., 2015). Networks trained on such large data sets are able to learn much more robust and complex features than those trained on smaller data sets. Therefore, instead of training our CNNs from scratch, we used pre-trained networks as a starting point, a procedure known as *transfer learning* (Yosinski, et al., 2014).

We downloaded an implementation of ResNet50 (He, Zhang, Ren, & Sun, 2016) that was pre-trained to perform image classification on the ILSVRC data set (other popular network architectures were also considered but were found to not perform as well). To adapt this network for our own purposes, we removed its topmost layers and replaced them with a new set of untrained layers so that we could take advantage of the low-level features trained on big data, while still being able to learn high-level features relevant to our specific task. More specifically, we kept each layer up to the final pooling layer, and then used global average pooling to convert the activation of the pooling layer into a vector that could be used as input into a series of fully-connected layers. For each of these layers, dropout (Srivastava, et al., 2014) and batch normalization (Ioffe & Szegedy, 2015) were used to improve generalization and accelerate learning. The dropout rate was set to 0.5, and the batch normalization parameters were left at their default values. Rectified linear units (ReLU; Nair & Hinton, 2010) were used as the activation functions. These layers fed into a final output layer consisting of 8 linear units corresponding to the 8 MDS dimensions.

Training Procedure

The objective function we sought to minimize was the mean squared error (MSE) between the network's output and the MDS coordinates of the rocks in the training set. To artificially increase the size of the training set we performed data augmentation: training images were randomly flipped, rotated, cropped, and stretched/shrunk every time they were presented to the network.

Training took place in two steps. During the first step we kept the parameters of the pre-trained CNN fixed and only trained the parameters of the newly-added fully-connected layers. Kingma and Ba's (2014) "Adam" was used as the optimization algorithm for this step. All of Adam's parameters were left at their default values except for the learning rate. The model was trained until validation error stopped decreasing for at least 20 epochs, or for a maximum of 500 epochs. During the second step, all of the network's parameters were trained. Because the parameters in the early layers were expected to already be close to their optimal values, stochastic gradient descent with a low learning rate and high momentum (0.0001 and 0.9, respectively) was chosen as the optimization algorithm. The network was trained for 500 epochs in this step, but only the weights from the epoch with the lowest validation error were saved.

We repeated this training procedure several times, each time using different values of *hyperparameters* (free parameters not learned by the network), with the goal being to find the hyperparameter values that yielded the lowest validation error. We optimized the following hyperparameters: the number of hidden layers added to the base CNN, the number of units in each hidden layer, the training batch size, and the initial learning rate. The optimal values were found to be 2, 256, 90, and $10^{-2.22}$, respectively.

Networks with the same architectures and hyperparameters may converge to different minima in the error space if their parameters are initialized to different random values, and it has been shown that combining the outputs of multiple networks usually yields better results than using any individual network (Hansen & Salamon, 1990). Therefore, after finding the optimal hyperparameter values, we repeated our training procedure 9 more times to produce an ensemble of 10 CNNs. Final predictions were produced by averaging the output of all 10 networks.

This ensemble achieved MSE=1.298 and R^2=0.780 on the validation set. While promising, this is likely an overestimate of true generalization performance because the ensemble was fit to the validation set. Therefore, in the next section we consider the ensemble's performance on the test set to get an unbiased estimate of its generalization ability.

Generalization Performance

Figure 1 plots the actual MDS values of the rocks from the test set against the values predicted by the ensemble of CNNs, as well as the correlations between the MDS values and CNN predictions. To be clear, none of the networks' parameters or hyperparameters were manipulated to decrease error on the test set, so these are true predictions of unseen data. As can be seen, the correlation between the ensemble's predictions and the actual MDS values is very high for most of the dimensions. The CNNs perform the best on the lightness and chromaticity dimensions, which is unsurprising given that these dimensions reflect low-level color information. It is also probably unsurprising that the CNNs perform less well on the "shape" dimension, since Nosofsky et al. (2017b) were not able to develop a clear interpretation of this dimension and speculated that it is actually an amalgamation of several underlying psychological dimensions. What may be surprising is that the CNNs perform almost as poorly on the roughness dimension as the shape dimension, even though the former seems to have a clearer interpretation. Inspection of the rocks the CNNs mis-predict reveals that there are several rocks located on the smooth side of the MDS space that actually have bumpy or wavy textures that appear rougher than their MDS coordinates would suggest. This indicates that there may be noise in Nosofsky et al.'s (2017b) MDS solution, a point to which we return in the General Discussion.

Overall, the ensemble of CNNs yields MSE=1.355 and R^2=0.767 on the test set. The fact that the ensemble accounts for over 75% of the variance in both the validation and the test sets provides converging evidence that deep learning networks can be trained to automatically extract psychological representations from previously unseen images. Now that we have demonstrated that the CNNs are

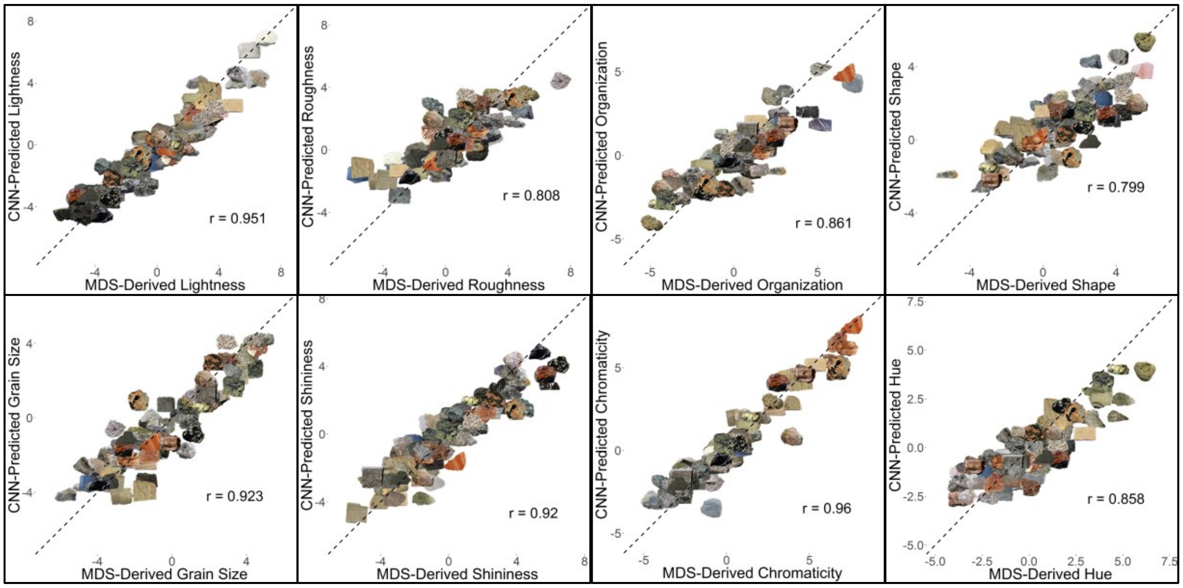

Figure 1: Scatterplot of MDS-derived dimensions against CNN predictions. r values indicate Pearson correlation coefficients.

capable of extracting MDS coordinates of novel stimuli, we turn to our next goal of using these representations to predict human categorization behavior.

Using Deep Learning-Derived Representations to Predict Human Categorization Behavior

In this section we describe a category learning experiment we conducted to test whether the CNN-derived representations could be used to predict human categorization behavior. We could not use images from the training or validation sets as stimuli in this experiment, because those images would not allow us to test how well the representations learned by the CNNs generalize to new rocks. We could have used the images from the test set, but we decided to do something more ambitious: we collected a completely new set of rocks. This approach allowed us to test whether the CNNs could predict human categorization using rocks that were not even in the same MDS solution that the networks were trained on.

Method

Participants The participants were 133 members of the Indiana University Bloomington community. Participants were compensated $10 with a possible $2 bonus for scoring at least 60% correct during the test phase of the experiment. Ultimately 8 participants were not able to achieve this criterion, and their data were excluded from further analyses.

Stimuli The stimuli were 120 images of rocks belonging to the same 30 categories used by Nosofsky et al. (2017b), although none of the individual images were repeated. Some of these new images were obtained through web searches, while others were photographs of rocks we took ourselves. Photoshopping procedures were used to remove backgrounds and idiosyncratic markings such as text labels from the images. Half of the images in each category were used as training items, and the other half were used as test items.

Procedure Each participant was randomly assigned to one of 3 conditions: igneous, metamorphic, or mixed. Participants in the igneous condition and metamorphic conditions were tasked with learning the 10 categories of igneous and metamorphic rocks, respectively, while participants in the mixed condition were presented with a mixture of igneous, metamorphic, and sedimentary categories (see Figure 3 for the specific categories used in each condition).

The experiment was divided into a training phase and a test phase. The training phase consisted of 6 blocks of trials. On each trial, participants were asked to categorize a single training item using the keyboard, and they were given feedback after entering their answer. Each training item was presented twice every block in random order. The test phase consisted of 4 blocks of trials. In this phase, each training and test item was presented once every block in random order, and no feedback was given for the test items. To keep participants engaged in the task, feedback was given for each training item once in the first two test blocks and once in the second two test blocks.

Model fitting

We fit a low-parameter version of the GCM to the categorization data, using the CNN-derived representations as input. For brevity we will refer to this model as GCM-CNN. In this model, the probability that item *i* is categorized into category *J* is found by summing the similarity of *i* to all exemplars of category *J* and then dividing by the summed similarity of *i* to all exemplars of all categories:

$$P(J|i) = \left(\sum_{j \in J} s_{ij}\right) / \sum_K \left(\sum_{k \in K} s_{ik}\right)$$

where s_{ij} is the similarity between item *i* and exemplar *j*. This similarity is given by

$$s_{ij} = \begin{cases} e^{-c_b d_{ij}}, \text{if } i \text{ and } j \text{ belong to different categories} \\ e^{-c_w d_{ij}}, \text{if } i \text{ and } j \text{ belong to the same category} \end{cases}$$

where d_{ij} is the CNN-derived Euclidean distance between item *i* and item *j*, and c_b and c_w are free parameters that determine the rate at which similarity declines with distance. Here, we allow different similarity gradients for between- and within-category comparisons because many categories of rocks have distinctive features that are not captured by the MDS representations but may nonetheless cause increased within-category similarity or decreased between-category similarity. For example, pumice can easily be recognized by its holey texture, but the presence of holes is not one of the MDS dimensions.

We fitted GCM-CNN to the categorization data by first calculating the proportion of correct categorization decisions in the test phase for all training items and all test items in each condition and each category of rock, averaged across all participants. We then searched for parameter values that minimized the MSE between the empirical observations and the model's predictions. We leave the modeling of individual participants and the time course of category learning during the training phase as topics for future research.

Results

Figure 2 displays the mean proportion of correct categorization decisions during the test phase as a function of condition and item type (training or test items). Inspection of this figure reveals that participants in all 3 conditions correctly categorized the training items nearly 100% of the time, indicating that errors on the test items were not simply due to failing to learn the training items. The figure also indicates that the categories in the mixed condition were somewhat easier to learn than those in the other conditions—test items in the mixed condition were correctly categorized nearly 80% of the time, while they were correctly categorized only about 60% of the time in the other conditions (chance performance is 10%). Most importantly, though, the figure shows GCM-CNN is able to quantitatively predict these patterns, achieving an MSE of 0.0005 and an R^2 of 0.97. The model does an excellent job of describing human categorization behavior at this coarse level of analysis.

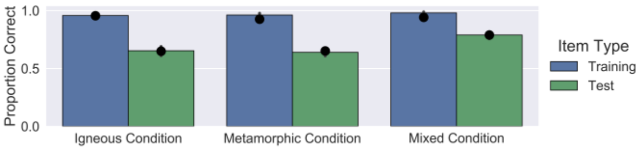

Figure 2: Mean proportion correct in the test phase as a function of condition and item type (training or test). Bar heights = observed data, error bars = 95% confidence intervals, dots = GCM-CNN predictions

Figure 3 presents a more fine-grained view of the data. This figure displays the mean proportion of correct categorization decisions for the test items as a function of condition and the individual categories of rocks (performance on the training items was near ceiling for every category). Inspection of this figure reveals that within each condition the categories varied in difficulty, and, generally speaking, GCM-CNN was able to predict which categories would be easy or hard. There are some notable exceptions, however. For instance, GCM-CNN under-predicts performance for pumice in both the igneous and mixed conditions. It seems that even with the inclusion of the c_w parameter, the model was not able to capture pumice's high amount of within-category similarity. Although our use of the c_w parameter provided a means of improving the shortcomings of the MDS representation, it may be necessary to train the CNNs to predict idiosyncratic features such as holes to fully capture human categorization behavior. Alternatively, it may be necessary to build prior knowledge into the model; some participants may have already been familiar with pumice stones because they are commonly used as exfoliants.

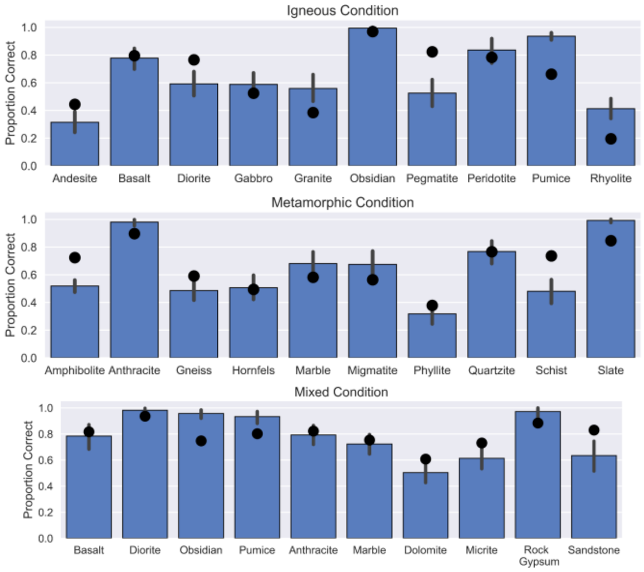

Figure 3: Mean proportion correct for test items as a function of condition and category of rock. Bar heights = observed data, error bars = 95% confidence intervals, dots = GCM-CNN predictions.

Inspection of Figure 3 also reveals performance differences for the same categories in different conditions. For example, performance for diorite was much higher in the mixed condition than in the igneous condition (likely because it could not be confused for the visually-similar granite in the mixed condition), and this pattern was correctly predicted by GCM-CNN. The model also correctly predicted that performance for anthracite was lower in the mixed condition than the metamorphic condition, likely because it was confused for obsidian—another category of dark, shiny rocks—in the mixed condition. However, GCM-CNN seems to have over-estimated the extent to which obsidian would be confused for anthracite, as it predicted a greater difference in performance for obsidian across the igneous and mixed conditions than was actually observed. Again, training the CNNs to predict more dimensions (obsidian can be distinguished from anthracite by its scalloped surfaces) or incorporating prior knowledge (obsidian is often referenced in popular culture) may lead to better predictions.

While GCM-CNN makes some mis-predictions regarding a few specific categories of rocks, we nonetheless find these results impressive overall, especially considering that only 2 free parameters were used, and the stimulus representations were machine-generated without any human input.

General Discussion

In this article we have shown that deep learning networks can not only learn psychological representations of complex natural stimuli, but that they can predict the representations of completely novel stimuli, and these representations can be used in combination with formal psychological models to predict human categorization behavior. These results provide promise that time- and resource-intensive MDS studies could be automated in the future, making possible more large-scale studies using natural stimuli. The results reported here should be regarded as a proof of concept and not as the absolute best results that our procedure could produce. For example, increasing the size of the dataset used for training is likely to improve the generalization power of the network. Likewise, it is almost certainly the case that more sophisticated versions of the CNNs and GCM could provide even more accurate predictions of the rocks' MDS coordinates and the human categorization data.

Perhaps even more importantly, future research will also explore ways to improve the MDS representations that serve as the training data for the CNNs. As alluded to earlier, there is likely noise in Nosofsky et al.'s (2017b) MDS solution because many entries in the 360x360 similarity matrix used to derive it were based on relatively few observations. Nosofsky et al. (2017b) outline several directions that might be pursued to develop still more accurate and comprehensive similarity-scaling solutions for the rock stimuli.

As noted in our introduction, our current proposal is meant to complement other recent approaches that have used CNNs to derive feature-space representations for naturalistic stimuli. The idea in these other approaches has been to use CNNs to directly predict classification or similarity data and then to use the representations learned by intermediate layers as candidates for the psychological representations of the stimuli. Our alternative proposal that we illustrated here was motivated by our concern that the extent to which the CNNs serve as adequate psychological models of human perception and learning remains unknown, whereas cognitive process models such as the GCM have undergone decades of testing. A fruitful direction of future research will involve systematic comparisons between these alternative approaches. Indeed, we believe it is important to test more systematically the utility of CNNs as models of human category learning and to compare their performance against GCM and other psychological models. The most successful models might then be retained and could be useful for guiding the search for effective methods of teaching categories. For example, simulation of the models could allow for an automated search of which training examples people should be shown to optimize their category learning and generalization (e.g., Khajah et al., 2014; Markant & Gureckis, 2014; Mathy & Feldman, 2016; Nosofsky et al., 2018; Patil et al., 2014).

Acknowledgments

This research was supported by NSF grant 1534014 (EHR Core Research) to Robert Nosofsky.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Devin, M. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

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.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence, 12(10), 993-1001.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. Paper presented at the International Conference on Machine Learning.

Khajah, M. M., Lindsey, R. V., & Mozer, M. C. (2014). Maximizing students' retention via spaced review: Practical guidance from computational models of memory. Topics in cognitive science, 6(1), 157-169.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep Neural Networks Predict Category Typicality Ratings for Images. Paper presented at the CogSci.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. Journal of Experimental Psychology: General, 143(1), 94.

Mathy, F., & Feldman, J. (2016). The influence of presentation order on category transfer. Experimental psychology, 63(1), 59-69.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. Paper presented at the Proceedings of the 27th international conference on machine learning (ICML-10).

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

Nosofsky, R. M., Sanders, C. A., Gerdom, A., Douglas, B. J., & McDaniel, M. A. (2017a). 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. (2018). 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. (2017b). Toward the development of a feature-space representation for a complex natural category domain. Behavior Research Methods, 1-27.

Patil, K., Zhu, X., Kopec, L., & Love, B. (2014). Optimal teaching for limited-capacity human learners. In Advances in Neural Information Processing Systems (NIPS), 2014.

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

Pothos, E. M., & Wills, A. J. (2011). Formal approaches in categorization: Cambridge University Press.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. Meyer & S. Kornlum (Eds.) Attention and Performance XIV. MIT Press.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Bernstein, M. (2015). Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211-252.

Shepard, R.N. (1980). Multidimensional scaling, tree-fitting, and clustering. Science, 210(4468), 390-398.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929-1958.

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. Journal of Memory and Language, 42(1), 51-73.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? Paper presented at the Advances in neural information processing systems.