Computational Brain & Behavior (2020) 3:229–251
https://doi.org/10.1007/s42113-020-00073-z

ORIGINAL PAPER

Training Deep Networks to Construct a Psychological Feature Space for a Natural-Object Category Domain

Craig A. Sanders[1] · Robert M. Nosofsky[1]

Published online: 14 January 2020

Abstract
Many successful formal models of human categorization have been developed, but these models have been tested almost exclusively using artificial categories, because deriving psychological representations of large sets of natural stimuli using traditional methods such as multidimensional scaling (MDS) has been an intractable task. Here, we propose a novel integration in which MDS representations are used to train deep convolutional neural networks (CNNs) to automatically derive psychological representations for unlimited numbers of natural stimuli. In an example application, we train an ensemble of CNNs to produce the MDS coordinates of images of rocks, and we show that the ensemble can predict the MDS coordinates of new sets of rocks, even those not part of the original MDS space. We then show that the CNN-predicted MDS representations, unlike off-the-shelf CNN representations, can be used in conjunction with a formal psychological model to predict human categorization behavior. We further show that the CNNs can be trained to produce additional dimensions that extend the original MDS space and provide even better model fits to human category-learning data. Our integrated method provides a promising approach that can be instrumental in allowing researchers to extend traditional psychological-scaling and category-learning models to the complex, high-dimensional domains that exist in the natural world.

Keywords Deep learning · Categorization · Computational modeling · Multidimensional scaling · Psychological representations

Introduction

Cognitive psychologists have proposed many formal models of human categorization (see Pothos and Wills 2011 for a review). These models have been successful at quantitatively predicting human behavior, but they have been tested almost exclusively using artificial categories composed of simple stimuli with small numbers of perceptual features. The use of such stimuli is convenient for modeling purposes because it is straightforward to derive psychological representations that can be used as input to the models (e.g., geometric forms can be represented in terms of shape, size, and color). Natural stimuli, on the other hand, may be composed of large numbers of complex psychological dimensions that cannot be so easily described or quantified. In addition, whereas research involving learning of artificial category structures typically involves the use of categories composed of relatively small numbers of items, categories in the natural world are composed of limitless numbers of items.

In the present work, we present initial research aimed at the development of a method that makes tractable the derivation of high-dimensional psychological scaling solutions for unlimited numbers of stimuli from complex, natural-category domains. The method involves a novel integration of traditional psychological scaling techniques and deep-learning networks. As described below, we illustrate the method in the domain of rock classification in the geologic sciences, although we believe that the general method should be applicable across wide varieties of naturalistic domains. Hence, the proposed method has the potential to significantly advance the testing and domains of application of computational models of cognition and behavior.

In recent work, Nosofsky, Sanders, and colleagues have applied categorization models to the problem of rock classification in the geological sciences (Nosofsky et al. 2017; Nosofsky et al. 2018a, 2018b; Nosofsky et al. 2018c; Nosofsky et al. 2019b). As seems to be true of almost all natural categories (Barsalou 1985; Rosch 1973), major types

✉ Robert M. Nosofsky
nosofsky@indiana.edu

[1] Department of Psychological and Brain Sciences, Indiana University Bloomington, 1101 E. Tenth Street, Bloomington, IN 47405, USA

of rock categories appear to have graded structures, with prototypical members at their centers, but with numerous less typical members as well. Thus, individual samples of the same type of rock often display remarkable within-category variability. In addition, the boundary lines separating contrasting rock categories from one another are often fuzzy, and the category distributions can sometimes even overlap. Finally, the rock categories are embedded in complex, high-dimensional feature spaces, and correct classification requires integrating information across these multiple dimensions. In the senses described above, rock categorization appears to be both a challenging and representative example of the forms of category learning that may operate in the natural world. Another advantage of conducting research in this domain is that relatively few people have detailed prior knowledge of the structure of rock categories in the geologic sciences; hence, the training history of category learning in this domain can be placed under careful experimental control in the laboratory.

To apply innumerable formal models of human classification learning in this and other natural-category domains, one needs to derive psychological representations of the stimuli, which are used as input to the models. In their initial studies, Nosofsky et al. (2018b, 2018c) used traditional multidimensional scaling (MDS; Lee 2001; Shepard 1980) procedures to derive these psychological representations. In brief, in typical MDS procedures, similarity judgments are collected for pairs of stimuli, and then the stimuli are placed in a feature space such that similar items are close together and dissimilar items are far apart. The dimensions that result from application of the procedure can then be interpreted and used as inputs to categorization models (for related examples of the procedure involving both naturalistic and semantic stimuli, see, e.g., Jones and Goldstone 2013; Roads and Mozer 2017; Voorspoels et al. 2008). Nosofsky et al. (2018c) conducted MDS analyses using a collection of 360 rocks from 30 major categories and found that the derived dimensions had sensible psychological interpretations, such as lightness of color and average grain size. Moreover, when used in combination with a well-known formal model of human categorization, these dimensions could be used to provide reasonably good quantitative accounts of performance in a variety of different category-learning experiments involving the rock stimuli (Nosofsky et al. 2018a, 2018b, 2019a).

Despite these initial successes, the MDS approach has some significant limitations. One of the most important limitations is a practical one: Deriving MDS representations of large numbers of stimuli requires an enormous amount of data. As the number of stimuli n grows larger, the use of the traditional psychological-scaling techniques for deriving MDS representations becomes intractable. In general, for n stimuli, there are $n(n-1)/2$ data cells in the lower triangle of a symmetric similarity-judgment matrix. Nosofsky et al.'s studies involving the 360 rocks therefore required obtaining data to fill 64,620 such data cells—and to obtain reliable data, numerous observations are required for each cell of the matrix. (Collecting this much data was actually so time- and resource-prohibitive, that the MDS space of Nosofsky et al. 2018c was ultimately derived from a similarity matrix where most cells were based on only one or two observations, and many cells were left completely empty.) If $n = 1000$, the number of cells rises to 499,500 and so on. Ultimately, a researcher may be interested in embedding an essentially unlimited number of objects from natural-category domains in high-dimensional scaling solutions, so any hope of using the traditional psychological-scaling method must be abandoned, even if one applies efficient adaptive routines to the collection of the similarity data (e.g., Roads and Mozer 2019; Tamuz et al. 2011). Although the specific example discussed above involved rock classification and similarity, it is clear that the same problem exists regardless of the natural-category domain under study.

The aim of the present work is to initiate the development of a technique that allows for the derivation of high-dimensional scaling solutions for unlimited numbers of natural-object stimuli. As explained in detail below, the idea is to combine the use of traditional psychological-scaling methods with the use of modern deep-learning technology and convolutional neural networks (CNNs, e.g., LeCun et al. 2015). In recent work, other researchers have also made use of deep-learning networks as an approach to deriving psychological feature representations for natural objects; as will be seen, however, our proposed approach is significantly different in its underlying spirit and may have some major advantages.

Perhaps the major current approach to deriving feature representations for large numbers of naturalistic stimuli is to start with "off-the-shelf" deep-learning CNNs, which have been trained to classify thousands of images from natural categories. Such CNNs learn representations of data in a hierarchical fashion inspired by the visual cortex and have been shown to spontaneously extract fundamental characteristics associated with perceptual and cognitive processing of natural objects (e.g., Nasr et al. 2019). These representations have been show to generalize to a wide variety of new computer vision tasks (Razavian et al. 2014). Such findings motivate the idea of treating the hidden-layer activations of the CNNs as candidates for the underlying psychological representations of the stimuli. Researchers have found, for example, that, once one makes allowance for certain mathematical transformations, CNN hidden-layer activations can be used to predict people's typicality ratings for objects from natural categories (Lake et al. 2015) and similarity judgments for natural objects (Peterson et al. 2018), as well as patterns of neural activity related to object categorization (e.g., Bashivan et al. 2019; Guest and Love 2017; Khaligh-Razavi and Kriegeskorte 2014; Yamins et al. 2014). These findings suggest that off-the-shelf CNNs may provide a ready source of representations that could be used as input to models of human categorization, an idea pursued with some success, for example, by Battleday et al. (2017, 2019) and Holmes et al. (2019).

Despite these promising results, however, there are reasons to be skeptical of the extent to which off-the-shelf CNN representations really mirror psychological representations. Some researchers have found strong qualitative differences between CNNs' and people's responses in visual search and categorization tasks (e.g., Eckstein et al. 2017; Geirhos et al. 2017; Jacobs and Bates 2019; Rajalingham et al. 2018), and it is well-known that CNNs may confidently classify two images that appear identical to the human eye into completely different categories (Szegedy et al. 2013). Such results suggest that CNNs and humans may make use of different sets of representations. Although these issues are a source of continuing debate (e.g., see Elsayed et al. 2018; Zhou and Firestone 2019), it is fair to say that the extent to which the learning and representational processes embedded in CNNs capture those of humans remains an open question.

Therefore, in our present work, rather than relying on the hidden-layer activations of CNNS as a source of psychological representations, we propose and begin the exploration of a complementary, alternative approach. The approach involves a novel integration of classic MDS methods and CNN technology. In our proposed approach, we do not treat CNNs as psychological models. Instead, we treat them as pure machine-learning models[1] and train them to produce the MDS coordinates of stimuli obtained in separate psychological scaling studies. Specifically, we are proposing a two-stage procedure. The first stage involves the typical hard work that is involved in using traditional psychological-scaling methods for deriving MDS representations for objects. Rather than scaling the entire domain of objects, however, in this first stage, one obtains a psychological scaling solution for only a representative subset of the domain of objects under study. The second stage then involves training CNNs to reproduce this psychologically derived scaling solution. If successful, then the method allows one to automate the embedding of an unlimited number of remaining objects from the domain into the derived psychological-scaling solution, thereby turning what was an intractable task into a manageable one.

The specific idea for the training of the CNNs is illustrated schematically in Fig. 1. We start by using CNNs that have been pretrained on thousands of natural images. Such networks provide powerful tools for extraction of fundamental elementary features that compose enormous varieties of natural images. We then attach to the final pooling layer of such CNNs a new set of fully connected layers to enable a form of *transfer learning* (see next section for details). Rather than training these networks to predict the category membership of visually presented natural objects (the standard approach that is currently used in the field), however, we instead train them to predict the *MDS coordinates* of those natural objects obtained from separate similarity-scaling studies (i.e., the similarity-scaling studies conducted during the first stage of the proposed method). The idea is that by using appropriate forms of training, the networks may generalize gracefully to produce the MDS coordinates of new stimuli as well. Thus, the networks could be used to automatically derive the psychological representations (MDS coordinates) of an unlimited number of objects from natural-category domains. In the examples in the present work, we train CNNs to take images of rocks as input and produce their psychological MDS coordinates as output. We then test whether the derived psychological representations can be used in combination with formal models to predict human category learning in independent experiments.

The use of connectionist networks as a means for extracting MDS representations has been proposed previously (e.g., Rumelhart and Todd 1993; Steyvers and Busey 2000)—again, however, there is a major conceptual distinction between our current approach and the past connectionist approaches. For example, Rumelhart and Todd (1993) showed that a shallow network could be used to extract representations of Morse codes. Their network took pairs of codes as input, and the codes were transformed into hidden-layer representations using two identical subnetworks. Similarities could then be computed from the learned representations using hard-coded computations, such as calculating the Euclidean distance between them. (Such networks are often called "Siamese" networks; see Bromley et al. 1994 and Chopra et al. 2005 for applications to signature and facial recognition, respectively.) Rumelhart and Todd found that after training such a network using similarity judgments collected from humans, the network's hidden layers represented the Morse codes in terms of their length and whether they were made of mostly dots or dashes. Precisely, the same dimensions for these stimuli have been uncovered using traditional MDS methods (e.g., see Kruskal and Wish 1978). Moreover, these representations generalized to stimuli from the same domain that the network was not trained on.

Although conceptually similar to our approach, Rumelhart and Todd's (1993) approach (as well as a related approach proposed by Steyvers and Busey 2000) differs in that they did not train the network to directly produce MDS dimensions; rather, they had the network indirectly learn psychological representations by training it to produce human similarity judgments. The Rumelhard–Todd and Steyvers–Busey approaches may initially seem more natural than the one we propose, because the MDS dimensions are not "ground truth" values, but are instead derived from the similarity judgments. Indeed, we leave open the possibility that the approaches initiated by researchers such as Rumelhart and Todd (1993) and Steyvers and Busey (2000) may prove to be preferable to our own.

[1] By a "pure machine-learning model," we mean that we are concerned only with the outputs that the CNNs produce, regardless of whether or not those outputs are achieved through human-like learning.

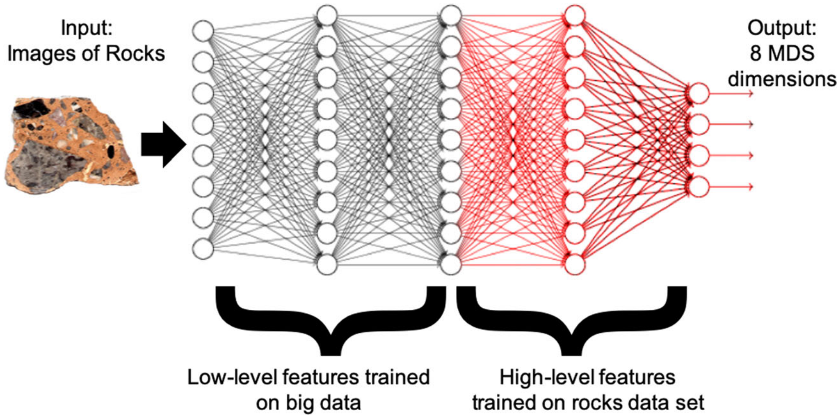


Fig. 1 Schematic illustration of our deep-learning approach. We start with a network pretrained to classify enormous numbers of images of objects from natural categories. We then append a new set of layers onto the network and retrain it to take images of rocks as input and produce their MDS coordinates as output

However, a potentially major advantage of our proposed approach is that it is directly extensible to other psychological dimensions not revealed by traditional similarity-scaling methods. For example, our rock classification studies have provided clear evidence that various psychological dimensions become highly salient in the context of the category-learning tasks themselves, but that these same dimensions may be completely ignored in the context of generic similarity-judgment tasks (Nosofsky et al. 2019a, 2019b). The reason seems to be that the dimensions are subtle ones, but they provide highly diagnostic cues for category membership (we provide specific examples later in our article). The phenomenon is closely related to the "feature-creation-and-discovery" ideas advanced by other researchers (e.g., Austerweil and Griffiths 2011, 2013; Schyns et al. 1998), in which new psychological features are created in the service of categorization. The important point here is that such dimensions are often not revealed by traditional similarity-scaling-based MDS studies; thus, it is not clear how approaches such as the ones suggested by Rumelhart and Todd (1993) and Steyvers and Busey (2000) would accommodate them. By contrast, as will be seen in our approach, it is straightforward to train CNNs to recognize these "missing" dimensions by manually adding them to the vectors of MDS-derived values that are used to train the networks.

In the remainder of this article, we describe how we trained CNNs to produce the MDS coordinates of the rocks from the data set of Nosofsky et al. (2018c), and we assess how well the CNNs are able to generalize these representations to both a held-out test set from within the original MDS space and to a set of completely new rocks outside the original MDS space. We then describe a new categorization experiment we conducted to assess whether the CNN-predicted MDS representations could be used in conjunction with a formal model of human categorization to predict the classification-learning data. We compare fits using the CNN-predicted MDS representations, the actual MDS representations, and off-the-shelf CNN representations as input to the model. Finally, we then explore if we can improve model fits to the behavioral data by supplementing the MDS space with additional sets of "missing" dimensions, and we assess the extent to which CNNs can learn these missing dimensions. Materials, code, and data from this article can be found in an online repository: https://osf.io/efjmq/.

Deep Learning Procedure

The goal of our deep learning procedure was to train CNNs that could take images of rocks as input and produce their MDS coordinates as output. Once the initial training is completed, the CCNs can be used to automatically generate psychological representations of infinite numbers of rocks. In this section, we provide an overview of the specific data set, CNN architecture, and training procedure that we used. Additional technical details can be found in Appendix 1.

Data Set

We made use of the data set of Nosofsky et al. (2018c), which consists of 360 images of rocks belonging to the three high-level categories of igneous, metamorphic, and sedimentary, with 10 subtypes within each high-level category, and 12 individual tokens within each subtype. The exact subtypes used in this data set can be found in Table 1. These subtypes are representative of those found in introductory college-level geology textbooks.

The data set also contains the values of each of the 360 rock-token images along 8 psychological dimensions, derived using MDS. The MDS procedures have been described extensively in previous articles (Nosofsky et al. 2018c, 2019a). In brief, participants provided similarity judgments between pairs of randomly selected rock images from the 360-item set. (Across two studies, there was a total of 198,555 pairwise similarity-judgment trials.) Initial MDS configurations were derived using nonmetric MDS procedures (Kruskal and Wish 1978) applied to the resulting 360 × 360 matrix of human similarity judgments among the rock images. Using the nonmetric solutions as starting configurations, a maximum-likelihood solution was derived using additional parameter-search routines. To improve interpretability, the maximum-

Table 1 Subtypes of igneous, metamorphic, and sedimentary rocks used in Nosofsky et al. (2018c) and the present work

Igneous	Metamorphic	Sedimentary
Andesite	Amphibolite	Bituminous coal
Basalt*	Anthracite*	Breccia
Diorite*	Gneiss	Chert
Gabbro	Hornfels	Conglomerate
Granite	Marble*	Dolomite*
Obsidian*	Migmatite	Micrite*
Pegmatite	Phyllite	Rock gypsum*
Peridotite	Quartzite	Rock salt
Pumice*	Schist	Sandstone*
Rhyolite	Slate	Shale

*Used in the mixed condition of the present experiment

likelihood solution was then rotated to achieve correspondence with sets of independent dimension-ratings data of the rock images obtained from other participants. The final solution and its dimensionality were chosen based on a combination of penalty-corrected maximum-likelihood fit (Lee 2001) and interpretability of the resulting dimensions.

The derived dimensions can be visualized interactively online: https://craasand.shinyapps.io/Rocks_Data_Explorer/. The first 6 dimensions are clearly interpretable in terms of lightness of color, average grain size, roughness, shininess, organization (rocks composed of organized layers vs. fragments haphazardly glued together), and chromaticity (warm/vivid colors vs. cool/dull colors). The interpretation of dimension 7 is not quite as clear-cut as the rest and likely reflects an amalgamation of several underlying psychological dimensions, but it seems to be related to shape (flat vs. spherical/cubical). Nosofsky et al. (2018c) initially considered dimension 8 to also have an ambiguous interpretation, but subsequent work found it can be well interpreted in terms of red versus green hue (Nosofsky et al. 2019a).

While the naïve approach would be to train and evaluate each network using all 360 images from the data set, 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 with varying *hyperparameters* (free parameters not learned by the network, such as the number of layers) 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 (see below and Appendix 1 for details). 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 stimuli. 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 and Training Procedure

One challenge we had to overcome was that the 360-rock set is quite small for a deep-learning data set. By comparison, image-classification networks are often trained 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 a pretrained implementation of ResNet50 (He et al. 2016) as a starting point.[2] This procedure is known as *transfer learning* (Yosinski et al. 2014). To adapt this network for our own purposes, we removed its output layer and replaced it with a new set of untrained fully connected 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. (The detailed procedure for deciding the structure of the appended fully connected layers is described in Appendix 1.) The final output of the network was 8 linear units corresponding to the 8 MDS dimensions.

We trained the network to minimize the mean squared error (MSE) between the network's output and the MDS coordinates of the rocks in the training set. Each of the dimensions was given equal weight in computing the MSE. (Note that the "importance" of each dimension is already reflected in the variance of the stimuli along that dimension in the original MDS solution that was derived from the similarity-judgment data.) 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. The images were scaled to a resolution of 224 × 224 pixels,[3] with the edges being cropped as necessary to make the images square without distorting their aspect ratios.

[2] We do not claim that there is anything special about the ResNet50 architecture; it simply yielded somewhat better model fits compared to the other architectures we tried, which included InceptionV3 (Szegedy et al. 2016) and VGG16/VGG19 (Simonyan and Zisserman 2014). For simplicity, we report the results from only the best-fitting network architecture among those that we tried. We emphasize as well that other more recently developed architectures such as InceptionResNet (Szegedy et al. 2017) or DenseNet (Huang et al. 2017) may yield even better results.

[3] This is the default image resolution assumed by ResNet50. Reducing the resolution to this size helps keep the training of the network computationally tractable, but it also obscures fine-grained details, which may have affected the networks' ability to learn some of the MDS dimensions.

The same network may converge to different minima in the error space if its 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 and Salamon 1990). Therefore, 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 (see Appendix 1 for example results involving the ensemble-based predictions). In the present case, 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, we must consider the ensemble's performance on the test set to get an unbiased estimate of its generalization ability.

Generalization to Rocks Within the Original MDS Space

Figure 2 displays scatterplots of the actual MDS values of the rocks from the test set against the values predicted by the ensemble of CNNs. *We emphasize here that these are "true" predictions without any human intervention or additional forms of parameter estimation*. High-quality versions of these plots and the exact MDS coordinates of each rock can be found in the online repository. 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 given that this dimension does not have a clear interpretation. Indeed, in our view, the fact that the networks are able to make even somewhat accurate predictions for this dimension is quite interesting and indicates that it does hold some meaning, even if that meaning is not immediately apparent to human observers.

What may be surprising about the ensemble's predictions is that the CNNs perform almost as poorly on the roughness dimensions as the shape dimension, even though the former seems to have a clearer interpretation. Inspection of the rocks that the CNNs mispredict reveals that there are several rocks located on the smooth side of the MDS space that actually have bumpy or wavy textures that seem rougher than their MDS coordinates would suggest. This indicates that there may be noise in the derived MDS space, which is unsurprising given that it is based upon an incomplete similarity matrix. We discuss possible directions for reducing noise in the MDS space 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 initial converging evidence that, if trained appropriately, deep learning networks can be used to automatically extract psychological representations for natural stimuli.

Generalization to Rocks Outside the Original MDS Space

We have emphasized that it is important to test the models on untrained stimuli to ensure that the models are generalizing to novel input and have not been overfitted to the training data. However, there is a sense in which our test set is not completely independent from the training or validation sets because all the sets came from the same MDS space. It is not clear that the same dimensions would emerge if the MDS analyses were conducted again using a new set of rocks, even if those rocks were sampled from the same categories used in the original set. If different MDS dimensions did emerge for different sets of rocks, then the CNNs would not actually be able to generalize to new stimuli, in spite of these results. Therefore, we conducted an MDS study using a new set of 120 rocks, belonging to the same categories as the 360-rock set, to assess whether the same dimensions would emerge again and whether the CNNs could generalize to this truly independent set of rocks.

Details of the new MDS study can be found in Appendix 2. In brief, we collected similarity ratings between each pair of the 120 new rocks, as well as independent ratings for each rock along the dimensions of lightness of color, average grain size, roughness, shininess, organization, and chromaticity. Then, following Nosofsky et al. (2018c, 2019b), we derived an 8-dimensional MDS space and rotated the first 6 dimensions of the space onto the dimension ratings to aid in interpretation. Figure 3 displays the rotated MDS space and Table 2 reports the correlations between the first 6 MDS dimensions and the direct dimension ratings. Figure 4 displays scatterplots between these MDS dimensions and the 8 predicted dimensions from the ensemble of CNNs. Again, these are true predictions without any additional forms of parameter estimation involved. Inspection of these figures reveals that as in the original 360-rock MDS space, dimensions 1, 2, 4, and 6 are interpretable in terms of lightness/darkness, average grain size, shininess, and chromaticity, respectively. These interpretations are corroborated by the high correlations between these MDS dimensions and the direct dimension ratings, as reported in Table 2. Furthermore, the correlations between these MDS dimensions and those predicted by the ensemble of CNNs are high, indicating that the ensemble is able to generalize to rocks that were not even included in the MDS space the networks were trained on. Dimension 8 can also again be interpreted in terms of hue—notice that there are many blue, purple, and red rocks at the bottom of Fig. 3 d, while there are more yellow, brown, and green rocks at the top. The red versus green contrast is not as pronounced in the

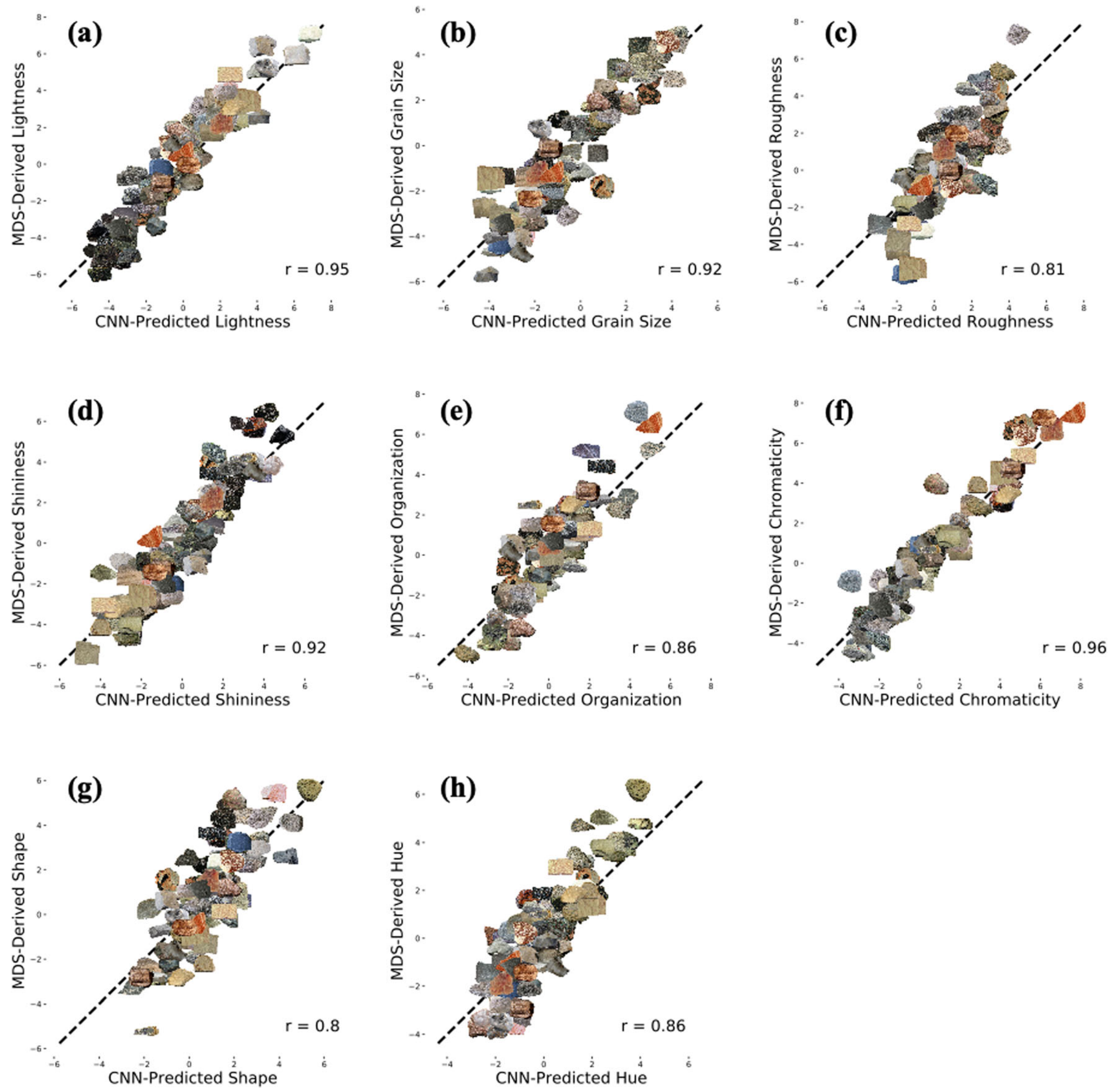


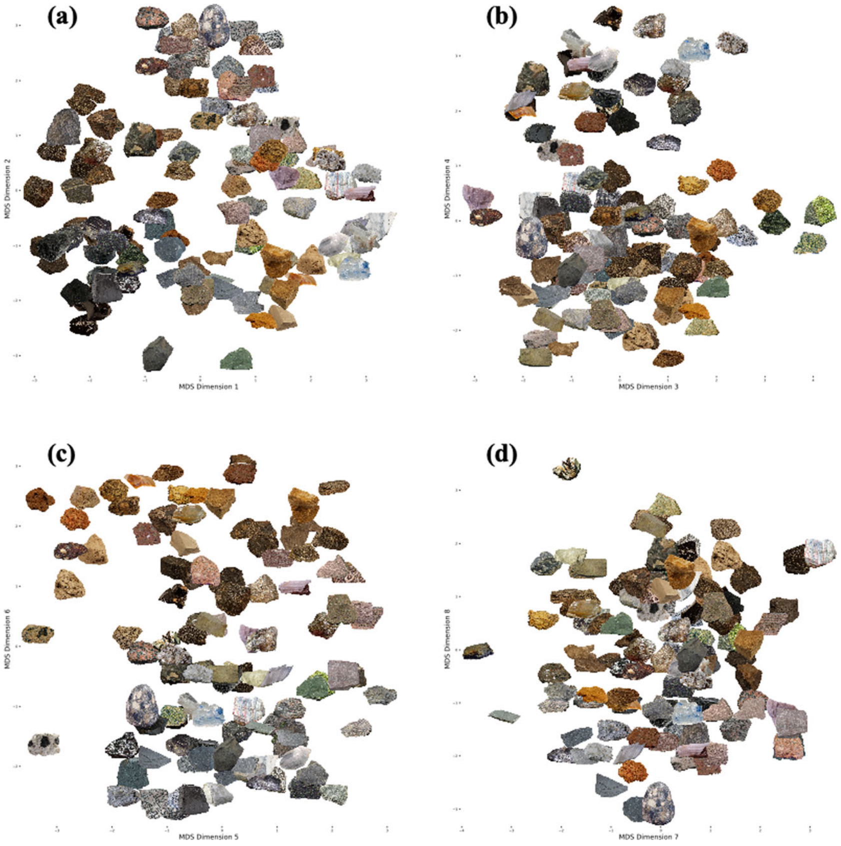
Fig. 2 **a**–**h** Scatterplots of CNN-predicted dimensions against MDS-derived dimensions for the test set. The *r* values indicate the Pearson correlation coefficients, and the dashed lines represent perfect prediction lines

120-rock set, however, which explains the somewhat lower correlation with the CNN-predicted hue values.

The interpretations of dimensions 3 and 5 in the 120-rock MDS space are not quite as clear-cut, however, as they were in the 360-rock MDS space. While it does seem to be generally true that rocks on the right side of Fig. 3 b are rougher than those on the left side, there are many exceptions, and the correlation between this MDS dimension and the direct roughness ratings was modest. Similar observations can be made for disorganized versus organized rocks in Fig. 3 c. Given that these dimensions failed to strongly replicate from the 360-rock MDS space, it is unsurprising that correlations between them and the CNN-predicted dimensions were relatively low. Note, however, that the CNN predictions still seem sensible. The rocks do seem to get gradually rougher as one moves from left to right in Fig. 4 c, and the rocks also seem to get gradually more organized as one moves from left to right in Fig. 4 e. Thus, it appears that the lower correlations may have more to do with differences in the derived MDS spaces across the two studies, rather than to any issues with the CNNs themselves.

Finally, given that dimension 7 in the 360-rock MDS space did not have a clear interpretation, we were surprised to see that a similar dimension nevertheless emerged again in this 120-rock MDS space. Notice that the rocks on the left side of Fig. 3 d tend to be flat, while the rocks on the right side tend to be more spherical or cubical, indicating that shape again influenced participants' similarity ratings. And while the correlation between this MDS dimension and the CNN-predicted dimension is relatively modest (Fig. 4g), the fact that the networks were able to generalize at all along this nebulous dimension is quite impressive. Moreover, the fact that this

Fig. 3 **a**–**d** Rotated MDS space for the 120-rock set



dimension emerged in both the 360-rock and 120-rock MDS spaces indicates that it really is psychologically meaningful, so future research will need to find a solid interpretation for it.

Now that we have demonstrated that, at least to a first approximation, the CNNs can generalize the MDS dimensions of Nosofsky et al. (2018c) to entirely new sets of rocks, we turn to our next main goal of using the CNN-predicted representations to predict human categorization behavior.

Categorization Experiment

This categorization experiment was conducted to compare different representations of the rocks on their ability to predict human categorization behavior when used in conjunction with a formal model of human category learning. The particular formal model that we use is Nosofsky's (1986, 2011) *generalized context model* (*GCM*). The GCM is a well-known model that has shown success in predicting human perceptual classification across numerous domains, including the present rock classification domain (e.g., Nosofsky et al. 2017, 2018a, 2018b, 2019a, 2019b). Moreover, it serves as a foundation for a number of other highly significant models of human category learning (e.g., Anderson 1991; Kruschke 1992; Love et al. 2004; Pothos and Bailey 2009; Vanpaemel and Storms 2008). Thus, it seemed like a reasonable starting point for use in the present investigation. We emphasize that the present experiment was not designed to provide tests between the GCM and other alternative models. Instead, the experiment and our use of the GCM are simply intended as tools for investigating the utility of the CNN-derived representations for predicting an independent set of human category-learning data. We expect that application of many closely related formal models of human classification would yield similar outcomes.

Table 2 Correlations between dimensions 1–6 of the rotated MDS space and the dimension ratings for the 120-rock set

Dimension	Correlation
1. Lightness of color	0.921
2. Average grain size	0.794
3. Roughness	0.542
4. Shininess	0.858
5. Organization	0.570
6. Chromaticity	0.798

There were three conditions in this experiment. Two of these conditions were conceptual replications of experiments conducted by Nosofsky et al. (2018b): the igneous condition, in which participants were tasked with learning the 10 subtypes of igneous rocks, and the mixed condition, in which participants were tasked with learning a mixture of igneous, metamorphic, and sedimentary rocks (see Table 1 for the

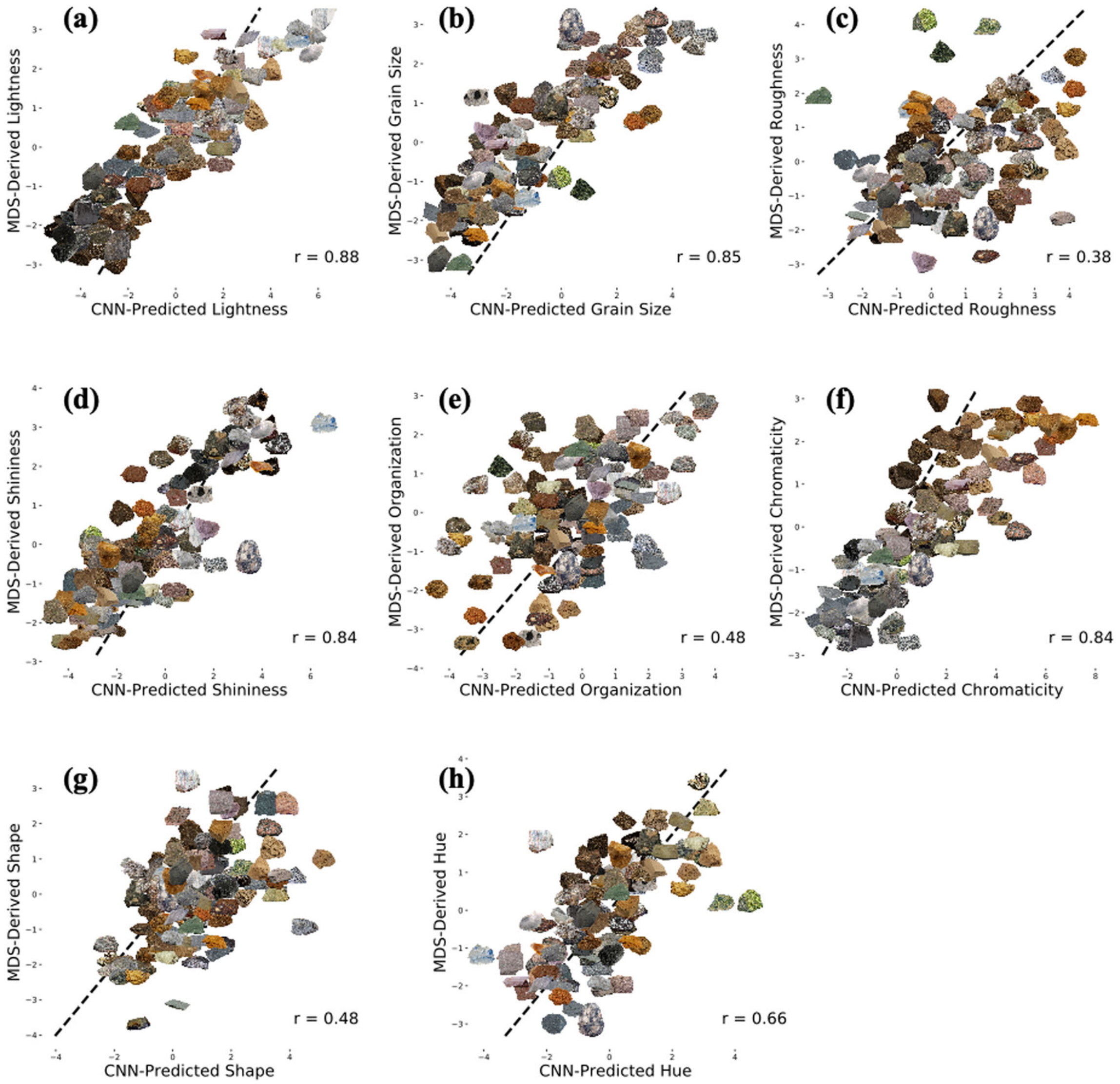


Fig. 4 **a**–**h** Scatterplots of CNN-predicted dimensions against MDS-derived dimensions for the 120-rock set. The *r* values indicate the Pearson correlation coefficients, and the dashed lines represent perfect prediction lines

specific subtypes used in the mixed condition). The third condition was the metamorphic condition, in which participants learned the 10 subtypes of metamorphic rocks. Importantly, in this design, certain subtypes of rocks appear across multiple conditions of category learning; thus, the design allowed us to test whether the derived representations could account for performance differences for the same subtypes across different conditions, based on changes in between-category similarity relations across the conditions.

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. There were 8 participants who did not achieve this criterion, and their data were excluded from further analyses, leaving 41 participants in the igneous and mixed conditions, and 43 in the metamorphic condition.

Stimuli

The stimuli were 120 images of rocks belonging to the same 30 subtypes used by Nosofsky et al. (2018c), although none of the individual images were repeated. There were 4 individual tokens in each subtype. Some of these images were obtained through web searches, while others were taken from a study reported by Meagher et al. (2018) that involved comparisons of human category learning of images of rocks versus physical samples. Photoshopping procedures were used to remove backgrounds and idiosyncratic markings such as text labels from the images. Within each subtype of rock, the first two tokens were selected as training stimuli, and the second two tokens were selected as transfer stimuli. Because there were 10 subtypes in each condition, there was a total of 20 training stimuli in each condition and 20 novel items presented at time of test.

Procedure

Each participant was randomly assigned to one of the three conditions: igneous, metamorphic, or mixed. The experiment was divided into a training phase and a test phase. The training phase consisted of 6 blocks of 40 trials each. On each trial, participants were presented with a single training item and asked to categorize it using the computer keyboard. Participants were given feedback after entering their response. The feedback always told participants the correct answer (e.g., "Correct, Andesite!" or "Incorrect, Basalt!"). Each of the 20 training items was presented twice every block in random order. The test phase consisted of 4 blocks of 40 trials each. In this phase, each training and transfer item was presented once every block in random order, and no feedback was given for the transfer items. Following previous work (e.g., Nosofsky et al. 2018b, 2019b), 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.

Modeling the Categorization Data Using MDS and CNN Representations

Complete classification confusion matrices from each of the three conditions can be found in the online repository. The matrices report the total number of times each individual rock was classified into each of the 10 available categories in each condition, aggregated across all subjects. Our first goal was to assess whether the actual MDS representations and CNN-predicted MDS representations of the rocks could be used in conjunction with the GCM to predict the categorization data. We fitted a low-parameter version of the GCM to the three confusion matrices from the test phase of the experiment, using a maximum-likelihood criterion (see the online repository for best-fitting parameters and predicted confusion matrices for all reported models). GCM is an exemplar model that assumes that people store exemplars of categories in memory and that stimuli are categorized according to how similar they are to these exemplars. Formally, the GCM states that the probability that item i is categorized into category J is found by summing the similarity of i to all training exemplars of category J and then dividing by the summed similarity of i to all training exemplars of all categories:

$$P(J|i) = \frac{\left(\sum_{j\in J} s_{ij}\right)}{\sum_K \left(\sum_{k\in K} s_{ik}\right)} \quad (1)$$

where s_{ij} denotes the similarity of item i to exemplar j. Similarity is computed as an exponential decay function of distance in psychological space (Shepard 1987):

$$s_{ij} = e^{-cd_{ij}} \quad (2)$$

where d_{ij} is the Euclidean distance between item i and exemplar j, and c is a free sensitivity parameter that determines the rate at which similarity decreases with distance. The GCM often includes additional parameters that determine the attention weights for the psychological dimensions, response biases for each category, memory strengths associated with individually stored exemplars, and the degree to which responding is probabilistic versus deterministic (for details, see Nosofsky 2011). However, for the present study, we focus on this "basic" version of the model that only uses c as a free parameter. In our view, our focus on this low-parameter version of the model is sensible given that our primary goal is to directly assess the utility of the CNN-derived representations for predicting the category-learning data. In addition, previous work (e.g., Nosofsky et al. 2019a) has already indicated that, in the present types of rock category-learning experiments, extending the model with these additional free parameters leads to relatively minor improvements in fit—especially compared to our use of "supplemental dimensions" described in the final model-fitting section of our article.

We fitted the basic GCM using both the standard (similarity-judgment-derived) MDS representations and the CNN-derived MDS representations of the rocks as input. Model fit diagnostics can be found in Table 3, and scatterplots of the models' predictions[4] and the observed classification probabilities can be found in Fig. 5. In these scatterplots, open symbols indicate within-category classifications (e.g., the probability that andesite was correctly classified as andesite), whereas x's indicate between-category classifications (e.g., the probability that andesite was incorrectly classified as basalt). While the standard MDS representations provide an overall better fit to the data, both the standard MDS and CNN-derived MDS representations are able to account for around 90% of the

[4] Whereas earlier in the article, we focused on out-of-sample predictions for the deep networks, here we focus on within-sample predictions of the GCM to be consistent with our previous work. We describe the predictions as "within-sample" because a free parameter c is being estimated to fit the data. As described in a later section, we used the BIC statistic (Schwarz 1978) to address the issue of overfitting for cases of models involving different numbers of free parameters.

Table 3 Number of free parameters of each version of GCM and its best-fitting negative log-likelihood, BIC score, and R^2

Model (representations)	Free parameters	$-\ln(L)$	BIC	R^2
Basic (MDS)	1	4503	9016	0.897
Basic (CNN-predicted MDS)	1	5772	11,553	0.882
Basic (off-the-shelf ResNet50)	1	8977	17,964	0.781
Basic (transformed ResNet50)	1	8604	17,217	0.797
Basic (120-transformed ResNet50)	1	6465	12,939	0.861
Extended (MDS + supplemental)	14	3427	6994	0.936
Extended (CNN-predicted: MDS + supplemental)	14	4874	9888	0.906

variance in the classification confusion matrices, and both representations are able to capture some important qualitative trends. For example, both predict correctly that accuracy for diorite should be higher in the mixed condition compared to the igneous condition because diorite is confused with granite in the igneous condition. (Granite and diorite are both light-colored, coarse-grained rocks composed of interlocking crystals.) Similarly, both predict correctly that accuracy for anthracite should be lower in the mixed condition compared to the metamorphic condition because anthracite is confused for obsidian in the mixed condition. (Anthracite and obsidian are both shiny black rocks.) These results lend further promise to the idea that deep learning could be used to automate MDS studies in the future.

Given that we are applying models with only a single free parameter to an extremely rich data set, and given that we are

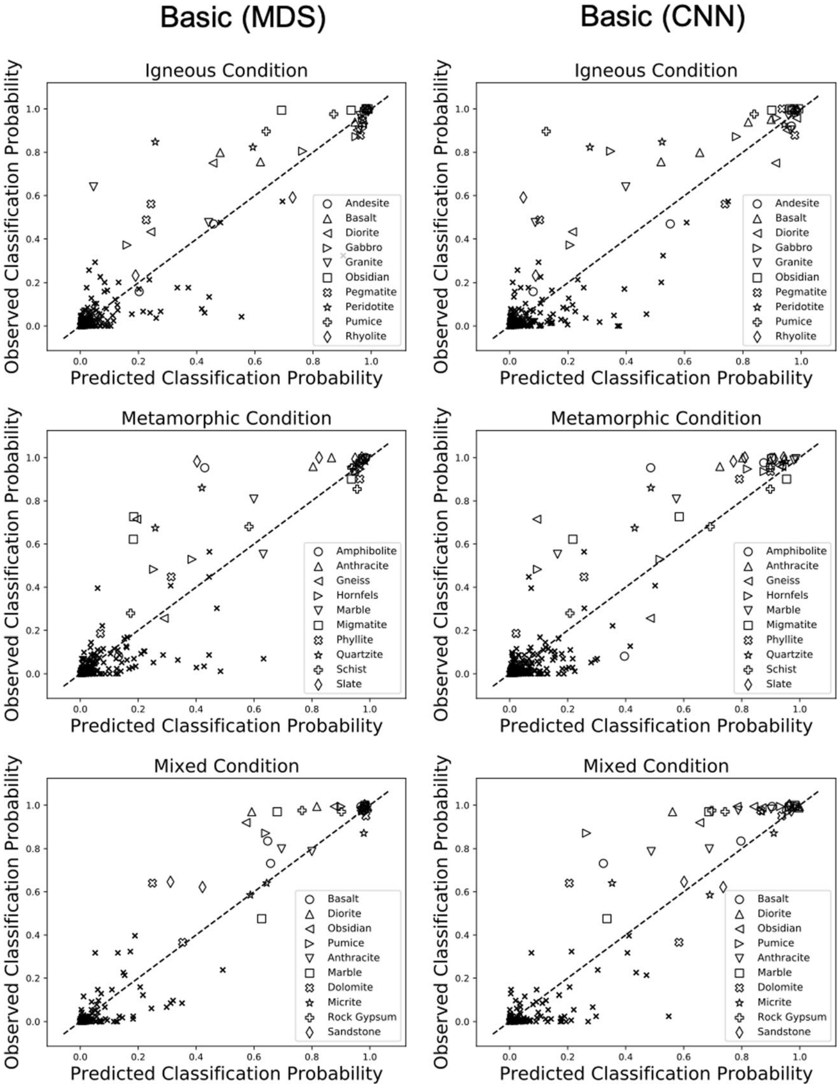


Fig. 5 Plots of GCM-predicted classification probabilities against observed classification probabilities. Left column: model predictions using the actual MDS representations as input. Right column: model predictions using the CNN-predicted representations as input. Rows: igneous, metamorphic, and mixed conditions

dealing with a complex, naturalistic-stimulus domain, these initial results appear to be promising. Nevertheless, the results also demonstrate a clear-cut limitation of the models: In particular, inspection of the scatterplots in Fig. 5 reveals that the models systematically underestimate the probability of correct classifications for many of the rocks, with numerous open-faced symbols lying above the perfect-prediction lines. This pattern was also discovered in a study reported by Nosofsky et al. (2019b). In a later section, we explore extensions of the MDS space and of the GCM to address this issue, but first we assess whether off-the-shelf CNN representations may be used in conjunction with the GCM to predict the human categorization behavior.

Modeling the Categorization Data Using Off-the-Shelf CNN Hidden-Layer Representations

As noted in the "Introduction," some researchers have used the hidden-layer activations of off-the-shelf CNNs to model psychological representations. This approach does not require any additional training of the networks, so it may seem preferable to our approach of training the CNNs to produce MDS coordinates. In this section, however, we show that the MDS-based representations are able to provide a much better account of the human categorization data than the off-the-shelf CNN hidden-layer representations, when used in conjunction with the GCM.[5]

To create off-the-shelf CNN representations of our rocks, we passed each rock image into a pretrained implementation of ResNet50 (other popular networks were also considered but were found to provide worse fits to the data) and extracted hidden-layer activations from the penultimate layer (an average pooling layer), creating a 2048-feature vector for each rock. We did this for both the 360-rock and 120-rock sets, and the resulting feature spaces can be found in the online repository. We then fitted the basic GCM model to the categorization data, using the ResNet50 feature vectors of the rocks from the 120-rock set as the input. Model fit diagnostics can be found in Table 3 (off-the-shelf ResNet50), and scatterplots of model predictions and observed classification probabilities can be found in Fig. 6. As can be seen, the standard MDS and CNN-predicted MDS representations provide much better fits to the data, suggesting significant limitations of the off-the-shelf CNN features as models of human representations.

It is possible, though, that some of the 2048 features are more important for classifying rocks than others, so a better fit could be found if the dimensions were appropriately weighted. Following a similar method as Peterson et al. (2018) and Battleday et al. (2017), we weighted the dimensions using a ridge regression model. The goal was to find a weighting of the dimensions that could predict the similarity relations between the rocks, and then use these weighted dimensions as input to the GCM. But because we ultimately want to develop an automated method for deriving psychological representations that does not require collecting additional similarity judgments, it did not seem appropriate to fit the ridge

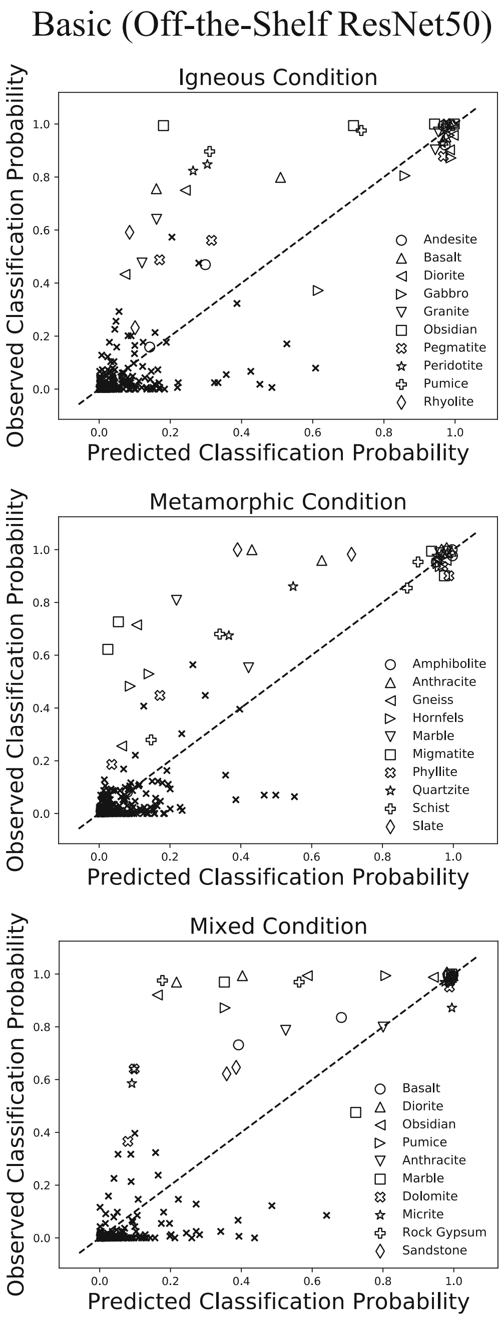


Fig. 6 Plots of GCM-predicted classification probabilities against observed classification probabilities, using off-the-shelf ResNet50 features as input

[5] As we discuss in detail in our "General Discussion," the hidden-layer-activation approach may still be viable if transfer learning were performed in which the CNNs were trained directly on the rock categories, with newly derived hidden-layer activations then being used as inputs to the psychological models. In this section, our focus is on only true "off-the-shelf" representations that do not require further training of the networks.

regression model to the similarity-judgment data of the same set of stimuli as used in the current classification experiment. Therefore, we first fitted the ridge regression weights to the similarity judgments from the 360-rock set of Nosofsky et al. (2018c), and we used the fitted values to transform the off-the-shelf ResNet50 representations of the current 120-rock set. This procedure also ensures that the ridge regression weights are not overfitted to the 120-rock set.

Formally, we converted the similarity judgments from the 360-rock set of Nosofsky et al. (2018c) into dissimilarity judgments by subtracting them from 10 (making 1 indicate least dissimilar and 9 indicate most dissimilar). The model's predicted dissimilarity between rock *i* and rock *j* is computed as the weighted Euclidean distance between their feature vectors[6]:

$$D_{ij}^{\text{pred}} = \sqrt{\sum_{m=1}^{2048} w_m \left(x_{im} - x_{jm}\right)^2} \qquad (3)$$

where $w_m > 0$ is the weight for dimension *m*, and x_{im} is the value of rock *i* on dimension *m* of the off-the-shelf CNN feature space. The objective function is minimization of the regularized squared error between the observed and predicted dissimilarities:

$$\Sigma_i \Sigma_j \left(D_{ij}^{obs} - D_{ij}^{\text{pred}}\right)^2 + \lambda \sum_{m=1}^{M} w_m^2 \qquad (4)$$

where λ is a regularization parameter. The latter term in this equation guards against overfitting by penalizing the model for putting too much weight on any one individual dimension.

We fitted this ridge regression model to the dissimilarity judgments from the 360-rock set of Nosofsky et al. (2018c). We first used 5-fold cross-validation to find the λ value that would yield the greatest generalization to the held-out set, and then we trained the model on the entire set of dissimilarity judgments to derive the *w* values. The optimal λ and *w* values can be found in the online repository. It is interesting to note that only 375 of the 2048 regression weights were nonzero, indicating that over 80% of ResNet50's features were irrelevant for predicting the similarity ratings.

We transformed the ResNet50 feature vectors for the 120-rock set by multiplying each dimension by its associated w_m; the transformed feature vectors can be found in the online repository. Finally, we fitted the basic GCM model again using these transformed feature vectors as input. Model fit diagnostics can be found in Table 3 (transformed ResNet50), and scatterplots of model predictions and observed classification probabilities can be found in the left column of Fig. 7. As can be seen, transforming the ResNet50 representations provides only a modest improvement in model fit, and the standard MDS and CNN-predicted MDS representations continue to provide a much better account of the data.

In a final analysis, we decided to give the ridge regression approach greater flexibility by fitting the ridge regression model *directly* to the similarity-judgment data of the 120-rock set, extracting the estimated feature weights, and then using those newly estimated weights for fitting the 120-rock classification data. (We reiterate our reluctance to follow this type of procedure, because the goal is to generate automated scaling solutions for the natural stimuli, rather than requiring that new sets of similarity-judgment data be collected for each new application.) Now, the model made use of more of the ResNet50 features, with 606 of the 2048 *w* values being nonzero. The key question, however, concerns the predictions of the 120-rock classification data. Using the off-the-shelf ResNet50 features, we again fitted the basic GCM to the classification data, except now using the weights that were derived by directly fitting the 120-rock similarity-judgment data. Model fit diagnostics can be found in Table 3 (120-transformed ResNet50), and scatterplots of model predictions and observed classification probabilities can be found in the right column of Fig. 7. As can be seen, the standard MDS and CNN-predicted MDS representations still provide much better fits to the human classification data than do the weighted ResNet50 features. These results suggest limitations to the approach of using simple linear transformations of off-the-shelf CNN hidden-layer activations as models of psychological representations of natural stimuli. We discuss possible reasons for these limitations as well as directions for alternative future approaches involving the use of CNN hidden-layer activations in our "General Discussion."

Extending the Models

Extending the MDS Space

As noted earlier in our article, even when it uses the standard MDS and CNN-predicted MDS representations, the basic GCM tends to underestimate correct classification probabilities for many of the rocks. Nosofsky et al. (2019a) showed that one reason for this limitation is that the MDS space underestimates within-category similarity because it is missing certain dimensions that are diagnostic of specific categories. The situation appears to arise because the dimensions are relatively subtle and tend not to be noticed in the context of generic similarity-judgment tasks; thus, they do not appear in the MDS solutions that are derived from the similarity-judgment data. However, because they are highly diagnostic for purposes of classification, the dimensions take on a good deal of salience in the context of the category-learning tasks themselves (cf. Austerweil and Griffiths 2013; Nosofsky 1986;

[6] Whereas we modeled *dissimilarities* using distances between feature vectors, Peterson et al. (2018) and Battleday et al. (2017) modeled *similarities* using dot-products between feature vectors. We found that our approach led to better model fits.

Schyns et al. 1998). Examples of some of these missing dimensions are illustrated in Fig. 8. From left to right, this figure shows examples of the rock types *andesite*, *pegmatite*, *obsidian*, *pumice*, and *slate*. Notice that the example of andesite has larger-sized fragments embedded in a more fine-grained groundmass, a pattern that geologists refer to as *porphyritic texture* (Tarbuck and Lutgens 2015). The example of pegmatite has a similar, but distinct pattern of banded dark crystals in its groundmass, a dimension we refer to as *pegmatitic texture*. The example of obsidian has a smooth sea-shell-shaped indent formed after a piece of the rock broke off, which is known as a *conchoidal fracture* (Tarbuck and Lutgens 2015). Finally, the example of pumice has many *holes*, and the example of slate has *physical la*yers.[7] Nosofsky et al. (2019a) found that extending the MDS space with these "supplemental" dimensions improved dramatically the GCM's ability to predict people's classification responses.[8] One virtue of the deep learning approach we have taken is that it is very modular: Rather than relying solely on the outputs produced by MDS solutions, the dimension-value training signals provided to the network can be imported from varied sources. Here we explored whether we could train CNNs to produce the supplemental dimensions noted above.

In their previous studies, Nosofsky et al. (2018c, 2019a) collected extensive ratings from participants of values of the items from the 360-rock set along a large number of individual dimensions, including the supplemental dimensions described above. From these data, we computed each rock's mean rating for the porphyritic texture, pegmatitic texture, and conchoidal fracture dimensions, as well as the proportion of participants who responded that the features "holes" and "physical layers" were present in each rock. We linearly rescaled the resulting mean-dimension ratings and feature-presence judgments for the 360 rocks to the range (− 5, 5) to make their scales comparable to the MDS dimensions. We then used our deep learning procedure to train an ensemble of CNNs to simultaneously predict the 360 rocks' mean ratings on the 5 supplemental dimensions as well as their values on the original 8 MDS dimensions. The same training, validation, and test sets were used as in our initial analyses. This ensemble achieved a MSE of 1.326 and R^2 of 0.737 on the validation set and a MSE of 1.404 and R^2 of 0.707 on the test set.

The ensemble's predictions for the test set are visualized in Fig. 9 (only the supplemental dimensions are shown to save space; predictions for the original MDS dimensions were comparable to those shown in Fig. 2). The CNNs' predictions for the supplemental dimensions are not quite as accurate as those for the MDS dimensions; a likely reason is that there are relatively few examples of rocks that clearly display the presence of positive values on these dimensions (i.e., the presence of holes, and so forth). Even so, the networks appear to do a reasonably good job of predicting the supplemental dimension values for these new rocks in the test set—again without any further intervention from the human user.

The question still remains how the CNNs would fare in automatically generating values on the supplemental dimensions of a completely new set of rocks that were rated by an independent set of participants. To find out, we collected ratings along the supplemental dimensions for the 120-rock set (see Appendix 2) and used the ensemble of CNNs to predict these ratings (without any further tuning of its parameters). Scatterplots of the predicted and observed ratings are shown in Fig. 10. Once again, the networks' predictions for the new set are not quite as accurate as for the original set, but they are at least in the right ballpark. Therefore, we now test whether these supplemental dimensions and the CNN-generated rating predictions can be used to improve the GCM's fits to the 120-rock classification data.

Extending the GCM

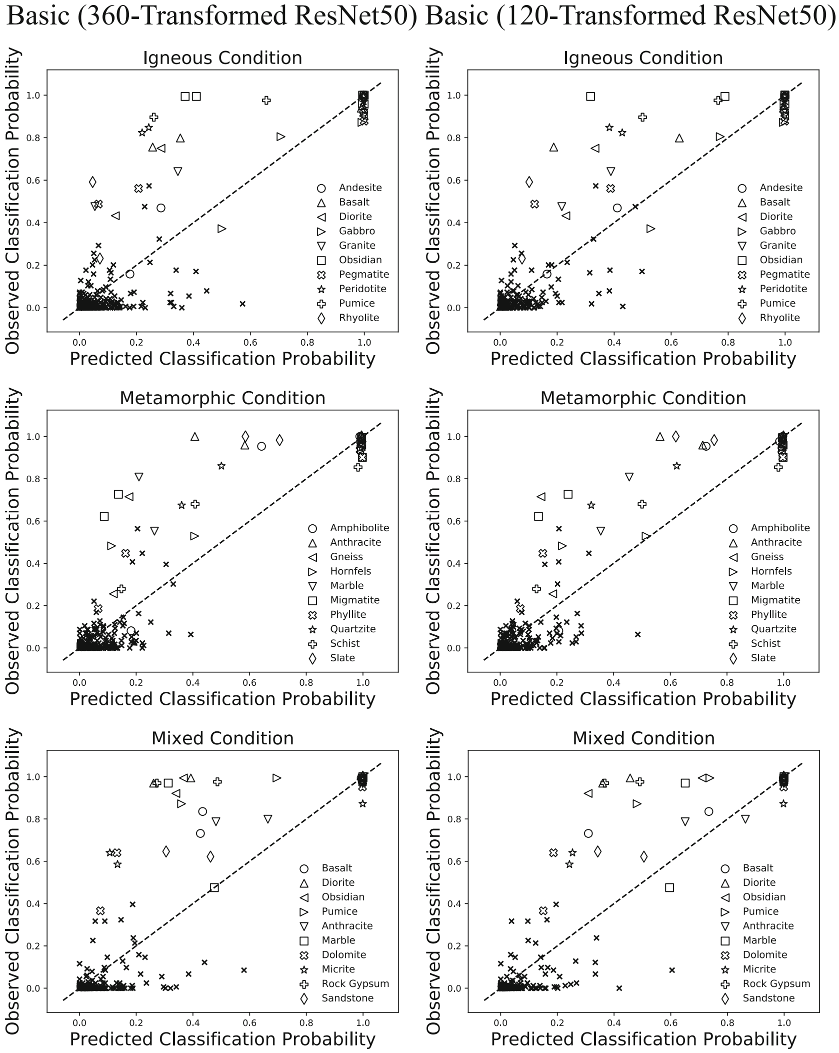
We refitted the GCM to the data obtained in our category-learning experiment by allowing the model to make reference to both the original MDS dimensions and also the supplemental dimensions. Because we do not know how the scales on the directly rated supplementary dimensions relate to those on the MDS-derived dimensions, we used an extended version of the GCM originally reported in Nosofsky et al. (2019a). Let r_{im} denote the average rating of rock i on rated dimension m. The psychological value of rock i on that dimension, r'_{im}, is given by the transformation

$$r'_{im} = \begin{cases} R_m + u(r_{im} - R_m)^p, & \text{if } r_{im} \geq R_m \\ R_m - v(R_m - r_{im})^q, & \text{if } r_{im} < R_m \end{cases} \qquad (5)$$

[7] The physical layers' dimension is partially captured by the "organization" MDS dimension, but the MDS space does not make a distinction between actual physical layers and stripes of different colors.

[8] Again, the aim of the present article is not to provide tests of the GCM against alternative models. Here, we simply use it as a tool for helping to evaluate the utility of alternative stimulus representations for predicting independent sets of classification-learning data. Nevertheless, one might argue that the need to expand the original MDS space with supplemental dimensions provides a challenge to the GCM, because typical applications make reference to only dimensions derived from independent sets of similarity-judgment data. In our view, this argument treats the exemplar-similarity model in a manner that is too constrained. People may classify objects based on their similarity to stored examples—whether the similarity comparisons are made in reference to "pre-existing" dimensions or to dimensions that are "discovered" in the service of categorization can be treated as a separate question. Yet another question is whether one needs to make use of the original similarity-judgment-derived MDS space at all: Why not simply create an *entire* researcher-defined set of features and collect direct ratings on all such features? Nosofsky et al. (2018b, 2018c) conducted extensive analyses to test such an approach, but found that the similarity-judgment-derived MDS space yielded far better accounts of both similarity-judgment data and independent sets of classification-learning data than did an approach that relied solely on participants' ratings of individual researcher-defined features. Understanding the detailed basis for those previous findings remains a topic for future research. Some possibilities are that it is difficult for participants to provide accurate ratings for individual dimensions when they are highly interacting with other dimensions, and that the psychological scales of the dimensions are highly nonlinear transforms of the direct dimension ratings provided by participants. MDS spaces derived from analysis of similarity-judgment data do not suffer from those problems.

Fig. 7 Plots of GCM-predicted classification probabilities against observed classification probabilities. Left column: model predictions using ResNet50 representations transformed by regressing onto the 360-rock similarities. Right column: model predictions using ResNet50 representations transformed by regressing onto the 120-rock similarities



where R_m is a “reference value” on the rated dimension, and u, v, p, and q are scaling constants (v can be held fixed at $v = 1$ without loss of generality). The parameters p and q allow for nonlinear relations between the psychological values and the direct ratings, and the reference value R_m allows for the shape of the nonlinear relation to vary with location on the rating scale. Nosofsky et al. (2019a) found that for values of the free parameters that tend to provide good fits to the data, this transformation behaves similarly to a step function: above the reference value R_m, a rock is considered to “possess” the relevant property, but the extent to which the rock is considered to have the property drops off sharply below that reference value.

Psychological distance in this extended GCM is given by

$$d_{ij} = \sqrt{\sum_{m=1}^{M}\left|x_{im}-x_{jm}\right|^2 + \sum_{m=1}^{M'} w_m\left|r'_{im}-r'_{jm}\right|^2} \qquad (6)$$

where x_{im} is the value of rock i on MDS dimension m, M is the number of dimensions in the original MDS space, M' is the number of supplemental dimensions, and w_m is the weight given to supplemental dimension m. As in the basic model, the distance d_{ij} is transformed to a similarity measure s_{ij} using Eq. (2), and the categorization probabilities are then computed using Eq. (1).

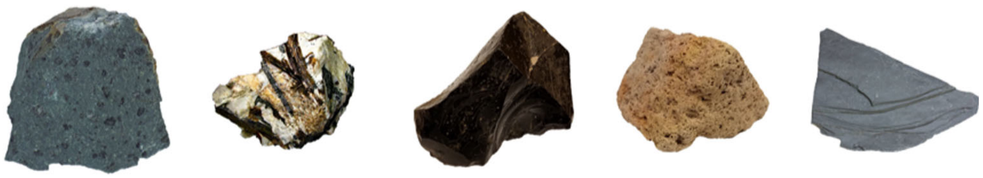


Fig. 8 Examples of rocks with properties not captured by the eight MDS dimensions. From left to right: andesite exhibits porphyritic texture, pegmatite exhibits “pegmatitic” texture, obsidian exhibits conchoidal fractures, pumice exhibits holes, and slate exhibits physical layers

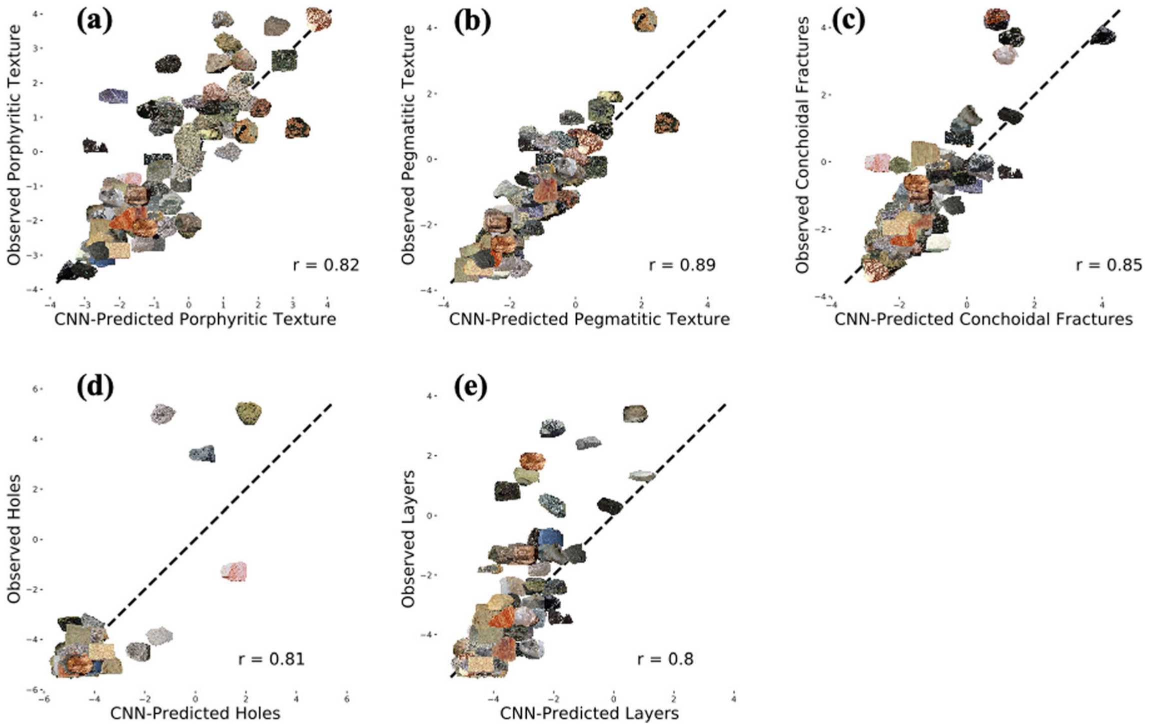


Fig. 9 **a**–**e** Scatterplots of CNN-predicted supplemental dimensions against human ratings for the test set from the 360-rocks study. The *r* values indicate the Pearson correlation coefficients, and the dashed lines represent perfect prediction lines

This extended GCM had 14 total free parameters: *c* from the basic model, the scaling constants *u*, *p*, and *q*, and a reference value, R_m, and weight, w_m, for each of the five supplementary dimensions. To compare the fits of this model to

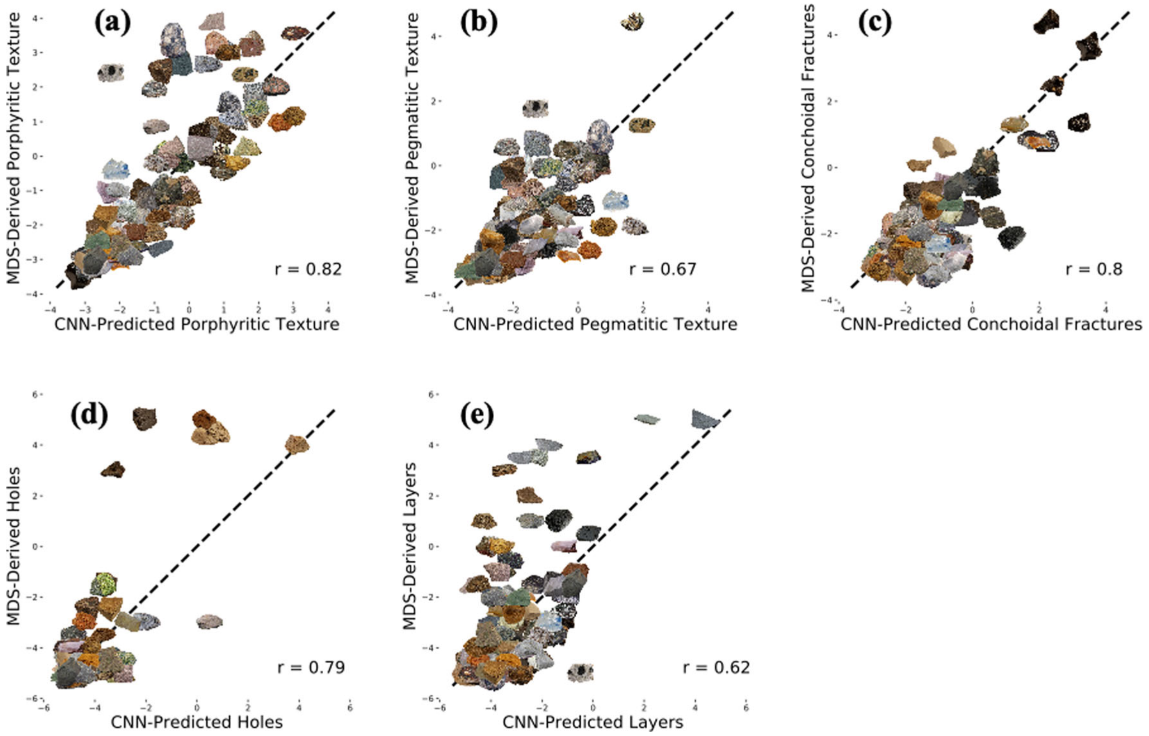


Fig. 10 **a**–**e** Scatterplots of CNN-predicted supplemental dimensions against human ratings for the 120-rock set. The *r* values indicate the Pearson correlation coefficients, and the dashed lines represent perfect prediction lines

those of the basic model, which only has one free parameter, we used the BIC statistic (Schwarz 1978), given by

$$BIC = -2lnL + Pln(N) \tag{7}$$

where L is the maximum-likelihood of the data, P is the number of free parameters in the model, and N is the total sample size of the data set. The latter term in Eq. (7) penalizes a model for having more free parameters. The model that yields the smallest BIC is considered to provide the most parsimonious account of the data.

We fitted two versions of this extended GCM to the categorization data: one that made reference to the standard MDS dimensions and the directly rated supplemental dimensions, and a second that made reference to the CNN-predicted values of the MDS and supplemental dimensions. Fit diagnostics for both versions of the extended model can be found in Table 3, and scatterplots of model predictions and observed classification probabilities can be found in Fig. 11. Despite having more free parameters, both extended models yield much better BIC scores than their corresponding basic models. Furthermore, inspection of Fig. 11 reveals that both extended models do indeed yield markedly improved accounts of the observed classification probabilities compared to the corresponding basic models. In particular, the extended models predict many more correct classifications for rocks possessing positive values on the supplemental dimensions, such as obsidian, pumice, and slate. There is still an overall tendency, however, for the models to underestimate the correct classification probabilities—most likely because there are diagnostic dimensions that are still missing from the MDS space. For example, we expect that adding a "stripes" dimension would improve the models' ability to account for the accuracy levels associated with the rock types gneiss and migmatite. As we have shown, it should be straightforward to train CNNs to produce these missing dimensions and any others that are found to be relevant for classifying the rocks.

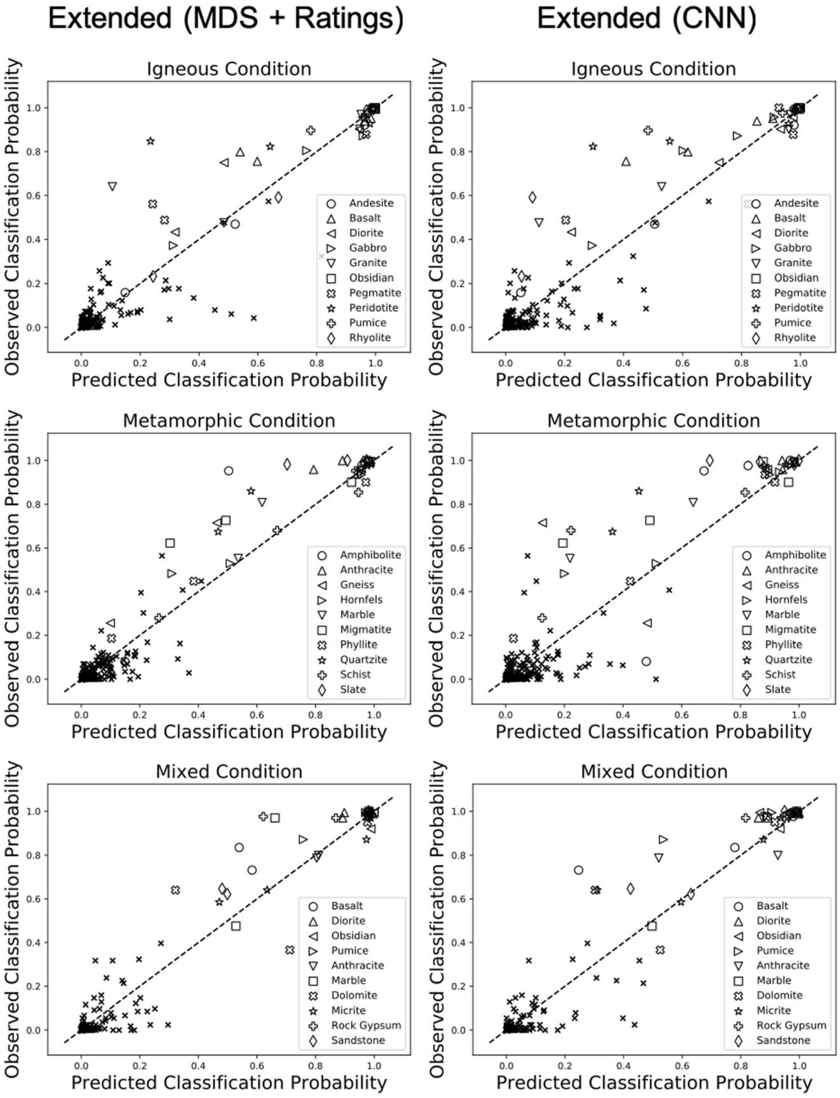


Fig. 11 Plots of GCM-predicted classification probabilities against observed classification probabilities. Left column: model predictions using actual MDS representations and ratings for the supplemental dimensions as input. Right column: model predictions using CNN-predicted MDS and supplemental dimensions. Rows: igneous, metamorphic, and mixed conditions

Despite these promising results involving the use of the supplementary dimensions, it is important to acknowledge that the version of the model that uses the CNNs to predict the supplementary-dimension values fares worse than the version that makes reference to the directly rated supplementary-dimension values (see Table 3). Thus, future work is needed to improve the CNNs' ability to automatically generate the scale values in the psychological feature space. We suggest routes for achieving this needed improvement in our "General Discussion."

General Discussion

Summary

In this research, we have taken promising steps toward the development of a deep-learning approach for embedding unlimited numbers of objects from natural-category domains in high-dimensional psychological spaces. The approach involves a novel integration of traditional MDS and deep-learning technology: In a first stage, traditional psychological scaling methods are used to derive MDS solutions for a representative subset of the domain of objects under study. In the second stage, the initially derived MDS solution is used to provide teaching signals to deep networks in order to directly train them to locate objects from the same domain in the derived psychological space. Admittedly, the approach does not remove the need for the painstaking work that is involved in deriving the starting MDS solution that is used for training the deep-learning networks. Crucially, however, once that starting MDS solution has been derived, the payoffs are potentially enormous: the approach allows for the automatic embedding in the psychological space of an unlimited number of additional objects from the relevant category domain. Furthermore, the same automatic embedding can be performed even if the objects reside in very high-dimensional, complex spaces. Thus, a goal that was formerly impossible to achieve—embedding unlimited numbers of real-world objects from natural categories in high-dimensional psychological spaces—is made tractable by the proposed approach.

In the present case, we considered only a single example target domain, namely rock classification in the geologic sciences. In our view, however, the same basic approach should be applicable regardless of the domain of inquiry. As we illustrated in the article, one would first derive an initial psychological space for a representative subset of objects from the domain by using a variety of complementary methods, including MDS analyses of similarity-judgment data and/or the use of direct dimension ratings. Once the initial MDS space is derived, it can then be used to train CNNs to generate representations for additional novel items from that domain that have not yet been scaled.[9] We should emphasize that although the examples provided in the present work involved only a relatively small set of cases for testing the ability of the networks to generalize to novel items, that restriction held only because we had a limited number of stimuli available for conducting the generalization tests. In practice, once the CNNs have been trained on the initial MDS solution, there is no limit on the number of new items from the domain that can be automatically scaled with the trained networks.

This automated MDS approach that we are developing could be instrumental in advancing cognitive theory and the testing of wide varieties of computational models of cognition and behavior. To take just one example, as noted at the outset of our article, most past research on computational modeling of human category learning has been restricted to the use of artificial category structures involving relatively small numbers of highly simplified, low-dimensional stimuli. Among the main reasons for that restriction is that the computational models make reference to a multidimensional feature space in which the to-be-classified objects are embedded. In particular, that multidimensional feature space serves as the input to the models (for extensive discussion, see Nosofsky 1992). To date, there have been no methods for deriving the high-dimensional feature space for large numbers of objects composing real-world natural categories. Thus, to the extent that the present approach is successful, rigorous quantitative tests of alternative computational models of human category learning can finally take place in domains of real wealth and significance, thereby allowing a deeper understanding of the nature of real-world human category learning to be achieved.

Indeed, once the high-dimensional feature space is derived, comparisons could even be conducted between well-known cognitive models of human category learning, such as exemplar and prototype models, and pure CNN models themselves. The cognitive models make clear-cut predictions about how patterns of generalization should vary across different training conditions, such as the precise sets of training examples that are experienced, whether there are differential payoffs for alternative categorization decisions, and so forth. At present, it is unclear how modern CNN models would respond to such experimental manipulations.

[9] Another specific example of deriving high-dimensional scaling solutions for complex, real-world categories is provided by the work of Getty, Swets, and their colleagues (Getty et al. 1988; Swets et al. 1991). Using a combination of MDS analyses of similarity judgments and direct ratings of individually specified dimensions, these investigators derived a 12-dimensional scaling solution for 24 instances of radiographs of benign versus malignant tumors in the domain of mammography. The derived dimensions corresponded to attributes such as roughness/smoothness of the border, the extent to which the tumor is invading neighboring tissue, the extent to which calcifications (small calcium deposits) are clustered, and so forth. Whereas Getty et al.'s MDS solution was limited to 24 instances of the radiographs of the benign and malignant tumors, with the present approach one could embed an unlimited number of such radiographs in the psychological scaling solution.

Limitations and Future Research

Although the results we reported in our article were promising, the predictions yielded by our use of CNN-derived MDS solutions were far from perfect, and use of the similarity-judgment-derived MDS solutions and directly rated supplemental dimensions allowed the GCM to achieve much better fits to the categorization data. Thus, an important direction for future research will be to improve the performance of the networks in producing the needed MDS solutions. Of course, one likely direction for such improvement will arise as researchers continue to enhance the technological sophistication of the networks themselves. Another way to move toward this goal is by providing the CNNs with better-quality training data. Regarding the specific cases described in this article, noise in the MDS space can be removed by collecting more similarity judgments and filling the missing entries in the 360 × 360 similarity matrix of Nosofsky et al. (2018c, 2019a). Furthermore, similarity judgments can be collected between rocks in the 360-rock set and the 120-rock set to create a shared 480-rock MDS space. Increasing the number of items in the MDS space may impose stronger constraints on where each item can be located, resulting in more accurate measurement of similarity relationships. Furthermore, embedding a larger number of objects in the MDS solution would create more training data for the CNNs, which would further improve their predictive power.

As a source of comparison, we also attempted to fit our category-learning data by using "off-the-shelf" CNN hidden-layer representations as input to the GCM (cf. Battleday et al. 2017, 2019). This approach fared much worse than the one we proposed, in which the CNN-trained MDS coordinates were used as input. Nevertheless, it is clear that continued exploration of the relationship between CNN hidden-layer representations and psychological representations will be a highly fruitful area of research. First, it is possible that more sophisticated transformations than the ones we used are necessary to align CNN and psychological representations. Second, the utility of CNN hidden-layer activations versus directly trained CNN MDS dimensions may vary with the target domain. The representations with the greatest utility may vary with the type of natural category being investigated or with the form of data of interest (e.g., behavioral choice data versus neural recordings). Third, it seems likely that the off-the-shelf CNNs we used were not sensitive to features relevant for rock classification simply because they were never directly trained on geoscience categories, and they could learn better representations through additional direct training on such objects. Pursuing this path is an extremely important one for future research, but will require extensive collection of a very large number of new images from the relevant rock categories to conduct such training (each of the rock categories in our current data set is composed of only 12 instances). We remark, however, that even if some of these suggested approaches to using hidden-layer activations are eventually shown to be successful, there may still be advantages to using MDS-based representations. For instance, deep-learning hidden-layer representations are often difficult to interpret, but uncovering semantically interpretable dimensions is one of the principal reasons for conducting an MDS analysis, and this interpretability can be important for advancing scientific theory.

Training networks to produce similarity judgments directly, as Rumelhart and Todd (1993) and Steyvers and Busey (2000) did, will also be an important direction for future research. Although we argued in the "Introduction" that this approach is limited because the networks cannot be easily trained to produce "missing" dimensions, there may be remedies to this problem. For example, such nets might be trained simultaneously on both similarity-judgment data and on classification data to discover a more complete set of psychologically relevant dimensions.

Although our discussion has focused on the shortcomings of the various feature spaces, there are also undoubtedly shortcomings in the GCM as a model of human categorization. Here, we used the GCM as a reasonable starting tool for conducting our investigations of the utility of the candidate feature spaces. It is possible, however, that alternative models such as clustering models (e.g., Love et al. 2004), Bayesian models (e.g., Anderson 1991; Sanborn et al. 2010), or rule-plus-exception models (e.g., Erickson and Kruschke 1998) could provide better fits to the data, and conclusions about the utility of the candidate feature spaces may vary with the specific model that is applied. In any case, each of these important computational models makes reference to a psychological feature space to generate its predictions. Our proposed approach to integrating MDS and deep-learning technology provides an important potential route to extending all of these computational models to account for performance in complex, high-dimensional category domains involving unlimited numbers of naturalistic stimuli.

Appendix 1

Details of Deep Learning Models

Our deep learning models were implemented using the Keras Python package (Chollet et al. 2015), the Scikit-learn Python package (Pedregosa et al. 2011), and the Tensorflow computational framework (Abadi et al. 2016). As mentioned in the main text, we took a transfer-learning approach (Yosinski et al. 2014), using a pretrained implementation of ResNet50 (He et al. 2016) as the base network. More specifically, we kept each layer from ResNet50 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 and Szegedy 2015) were used to improve generalization and accelerate learning. Rectified linear units (Nair and Hinton 2010) were used as the activation functions. The dropout rate was set to 0.5, and the hyperparameters for batch normalization were left at their default values. These layers fed into a final output layer consisting of 8 linear units corresponding to the 8 MDS dimensions.

We minimized the mean squared error (MSE) between the network's output and the MDS coordinates of the rocks in the training set, using Kingma and Ba's (2014) "Adam" as the optimization algorithm, with all of its hyperparameters left at their default values except for the learning rate. The network was trained until validation error stopped decreasing for at least 20 epochs, or for a maximum of 500 epochs. Only the newly added fully connected layers were trained at this stage. We used the hyperopt Python package (Bergstra et al. 2013) to optimize 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. This model achieved a MSE of 1.494 on the validation set. For comparison, the lowest validation error we could achieve without using transfer learning was 1.856.

To further reduce validation error, the transfer-learning network was trained for another 500 epochs, using a *fine-tuning* procedure (Yosinski et al. 2014). This time all layers were trained. Because the parameters in the base CNN 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. After fine-tuning, the network achieved a MSE of 1.330 on the validation set. We repeated this entire procedure 9 more times to produce an ensemble of 10 CNNs. Final predictions were produced by averaging the output of all 10 networks. Each network in the ensemble had the same hyperparameter values. Code for training this ensemble can be found in the online repository (https://osf.io/efjmq/). This ensemble achieved MSE = 1.298 on the validation set and MSE = 1.355 on the test set.

A reviewer of an earlier version of the article was interested in the extent to which there was variability across different runs of the network and the degree of improvement achieved through using the ensemble-based predictions. Unfortunately, we did not record the individual network fits in conducting the original versions of these massive deep-learning investigations. However, to provide a sense of the issue, we repeated the training procedures except using a smaller number of total training epochs (200) than used for the results reported in the main text. The MSEs and R^2s obtained for the validation and test sets for these reduced-training runs are reported for each individual network run and for the ensemble predictions in Appendix Table 4. As can be seen, the variability in fits across the individual network runs is relatively small, with a modest improvement in overall fit achieved by making using of the ensemble-based predictions.

Finally, to predict the supplementary dimensions, we created a new ensemble using the exact same procedure, but the networks were trained to predict both the 8 MDS dimensions and the 5 supplemental dimensions. The optimal parameter values this time were 3, 512, 30, and $10^{-2.05}$ for 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, respectively. This ensemble achieved a MSE of 1.326 on the validation set and 1.404 on the test set.

Appendix 2

Method for Collecting Similarity Judgments and Dimension Ratings

We closely followed the procedures for collecting similarity judgments and dimension ratings described in Nosofsky et al. (2018c). These data are available in the online repository. (https://osf.io/efjmq/).

Participants

The participants were 174 students from the Indiana University, Bloomington community. Data from 11 participants were removed because their responses had low correlations with the averaged responses. Some participants received credit toward a course requirement, while others received $12 as compensation. All participants reported normal or corrected-to-normal vision and no expertise in geology. Of these participants, 85 provided similarity judgments; 20 provided ratings for the lightness/darkness of color, average grain size, and smoothness/roughness dimensions; 20 provided ratings for the shininess, organization, and chromaticity dimensions; 20 provided ratings for the porphyritic texture, conchoidal fractures, holes, and layers dimensions; and 29 provided ratings for the pegmatitic texture dimension.

Stimuli

The stimuli were the 120 rock images used in the categorization experiment described in the main text.

Similarity-Judgments Procedure

Participants were shown pairs of rock pictures and were instructed to judge the similarity of the rocks on a scale from 1 (most dissimilar) to 9 (most similar). On each trial, two subtypes were randomly selected, and then one token was randomly selected as a representative within each subtype (the same token could not be selected twice when the subtypes were the same). One token was placed on the left side of the screen, and the other was placed on the right. The participants gave their judgment for the pair using the computer keyboard. This procedure was repeated for all 435 unique pairs of the 30 rock subtypes, as well as all 30 within-subtype comparisons, for a total of 465 trials. Participants first completed 5 practice trials to get a sense of the types of stimuli they would see. (Because we removed the data of 6 participants due to low correlations with the averaged data, the data from a total of 79 participants—a total of 36,735 similarity-judgment trials—were included in the MDS analysis.)

Dimension-Ratings Procedure

Participants gave ratings for one dimension at a time. First, instructions explaining the dimension and its rating scale were shown. Then, on each trial, participants were shown one of the 120 rocks and were asked to provide a rating on a 1–9 scale along the dimension, with the exceptions of the holes and layer dimensions. For these dimensions, participants indicated whether each rock had holes, layers, or neither (no rock had both). Responses were entered using the computer keyboard. To promote a consistent scale across participants for each dimension, the scale was shown at the bottom of the screen with labeled anchor pictures at the middle and extreme ends of the scale. See the online repository for each dimension's instructions and anchor pictures.

Table 4 Fit results from individual network runs and for the ensemble-based predictions for the MDS dimensions in the original 360-rock study

Network run	Validation set		Test set	
	MSE	R^2	MSE	R^2
1	1.408	0.760	1.548	0.734
2	1.450	0.754	1.603	0.726
3	1.455	0.752	1.538	0.738
4	1.454	0.751	1.554	0.736
5	1.430	0.756	1.551	0.738
6	1.414	0.759	1.552	0.737
7	1.464	0.752	1.617	0.723
8	1.439	0.752	1.577	0.731
9	1.470	0.748	1.579	0.730
10	1.432	0.758	1.586	0.731
Ensemble	1.387	0.763	1.509	0.742

Note. The results reported in this table are for reduced-training runs involving only 200 training epochs rather than for the full-training runs reported in the main text of the article

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: large-scale machine learning on heterogeneous distributed systems. *ArXiv Preprint ArXiv, 1603*, 04467.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*(3), 409.

Austerweil, J. L., & Griffiths, T. L. (2011). A rational model of the effects of distributional information on feature learning. *Cognitive Psychology, 63*(4), 173–209.

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review, 120*(4), 817–851.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 629.

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science, 364*(6439), eaav9436.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2017). Modeling human categorization of natural images using deep feature representations. *ArXiv:1711.04855 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/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.

Bergstra, J., Yamins, D., & Cox, D. (2013). Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures (pp. 115–123). Presented at the International Conference on Machine Learning.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1994). Signature verification using a "Siamese" time delay neural network. In *Advances in neural information processing systems* (pp. 737–744).

Chollet, F., et al. (2015). Keras.

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 539–546 vol. 1). https://doi.org/10.1109/CVPR.2005.202.

Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology, 27*(18), 2827–2832.e3. https://doi.org/10.1016/j.cub.2017.07.068.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology. General, 127*(2), 107–140. https://doi.org/10.1037//0096-3445.127.2.107.

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both human and computer vision. *arXiv preprint, arXiv, 1802.08195 10*.

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *ArXiv Preprint ArXiv, 1706*, 06969.

Getty, D. J., Pickett, R. M., D'Orsi, C. J., & Swets, J. A. (1988). Enhanced interpretation of diagnostic images. *Investigative Radiology, 23*(4), 240–252.

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

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(10), 993–1001. https://doi.org/10.1109/34.58871.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition (pp. 770–778). Presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Holmes, W. R., O'Daniels, P., & Trueblood, J. S. (2019). A joint deep neural network and evidence accumulation modeling approach to human decision-making with naturalistic images. Computational Brain & Behavior, 1–12.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700–4708).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift (pp. 448–456). Presented at the International Conference on Machine Learning.

Jacobs, R. A. & Bates, C. J. (2019). Comparing the visual representations and performance of human and deep neural networks. *Current Directions in Psychological Science, 28,* 34-39.

Jones, M., & Goldstone, R. L. (2013). The structure of integral dimensions: contrasting topological and Cartesian representations. *Journal of Experimental Psychology: Human Perception and Performance, 39*(1), 111–132.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology, 10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915.

Kingma, D., & Ba, J. (2014). Adam: a method for stochastic optimization. ArXiv Preprint ArXiv*:*1412.6980

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

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills: Sage.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. Presented at the CogSci.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539.

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology, 45*(1), 149–166.

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

Meagher, B. J., Cataldo, K., Douglas, B. J., McDaniel, M. A., & Nosofsky, R. M. (2018). Training of rock classifications: the use of computer images versus physical rock samples. *Journal of Geoscience Education, 66*(3), 221–230. https://doi.org/10.1080/10899995.2018.1465756.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines (pp. 807–814). Presented at the Proceedings of the 27th international conference on machine learning (ICML-10).

Nasr, K., Viswanathan, P., & Nieder, A. (2019). Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science Advances, 5*(5), eaav7903.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology. General, 115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39.

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

Nosofsky, R. M. (2011). The generalized context model: an exemplar model of classification. In Pothos, E. M. and Wills, A. J. (Eds.), *Formal approaches in categorization*, 18–39. Cambridge University Press.

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. https://doi.org/10.1177/0956797616675636.

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*(2), 129–135. https://doi.org/10.1177/0963721417740954.

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*(3), 328–353. https://doi.org/10.1037/xge0000369.

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*(2), 530–556. https://doi.org/10.3758/s13428-017-0884-8.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., Douglas, B. J. (2019a). Search for the missing dimensions: building a feature-space representation for a natural-science category domain. Computational Brain & Behavior, 1–21

Nosofsky, R. M., Sanders, C. A., Zhu, X., & McDaniel, M. A. (2019b). Model-guided search for optimal natural-science-category training exemplars: a work in progress. *Psychonomic Bulletin & Review, 26*(1), 48–76.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Dubourg, V. (2011). Scikit-learn: machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science, 42*(8), 2648–2669. https://doi.org/10.1111/cogs.12670.

Pothos, E. M., & Bailey, T. M. (2009). Predicting category intuitiveness with the rational model, the simplicity model, and the generalized context model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 1062.

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

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *BioRxiv, 240614*. https://doi.org/10.1101/240614.

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 806–813).

Roads, B. D., & Mozer, M. C. (2017). Improving human-machine cooperative classification via cognitive theories of similarity. *Cognitive Science, 41*(5), 1394–1411.

Roads, B. D., & Mozer, M. C. (2019). Obtaining psychological embeddings through joint kernel and metric learning. *Behavior Research Methods, 51*, 2180–2193. https://doi.org/10.3758/s13428-019-01285-3.

Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories. In *Cognitive development and acquisition of language* (pp. 111–144). Academic Press.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. Attention and performance XIV*:* synergies in experimental psychology, artificial intelligence, and cognitive neuroscience, 3–30.

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.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review, 117*(4), 1144–1167.

Sanders, C. A. (2018). Using deep learning to automatically extract psychological representations of complex natural stimuli. Unpublished Ph.D. dissertation, Indiana University.

Sanders, C. A., & Nosofsky, R. M. (2018). *Using deep learning representations of complex natural stimuli as input to psychological models of classification*. Madison: Proceedings of the 2018 Conference of the Cognitive Science Society.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences, 21*(1), 1–17.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210*(4468), 390–398. https://doi.org/10.1126/science.210.4468.390.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317–1323.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

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.

Steyvers, M., & Busey, T. (2000). Predicting similarity ratings to faces using physical descriptions. Computational, geometric, and process perspectives on facial cognition: contexts and challenges, 115–146.

Swets, J. A., Getty, D. J., Pickett, R. M., D'Orsi, C. J., Seltzer, S. E., & McNeil, B. J. (1991). Enhancing and evaluating diagnostic accuracy. *Medical Decision Making, 11*(1), 9–17.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818–2826).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *ArXiv:1312.6199 [Cs]*. Retrieved from http://arxiv.org/abs/1312.6199

Tamuz, O., Liu, C., Belongie, S., Shamir, O., & Kalai, A. T. (2011). Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*.

Tarbuck, E. J., & Lutgens, F. K. (2015). *Earth science* (14th ed.). Boston: Pearson.

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: the varying abstraction model of categorization. *Psychonomic Bulletin & Review, 15*(4), 732–749.

Voorspoels, W., Vanpaemel, W., & Storms, G. (2008). Exemplars and prototypes in natural language concepts: a typicality-based evaluation. *Psychonomic Bulletin & Review, 15*(3), 630–637.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, 111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? (pp. 3320–3328). Presented at the Advances in neural information processing systems.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications, 10*(1), 1334.

This article is based on a PhD dissertation submitted by the first author to Indiana University (Sanders 2018). A report of preliminary versions of some of the work reported in this article was published in the 2018 Proceedings of the Cognitive-Science Society (Sanders and Nosofsky 2018). This research was supported by NSF grant 1534014 (EHR Core Research) to Robert Nosofsky.