COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13226
© 2023 Cognitive Science Society LLC.
ISSN: 1551-6709 online
DOI: 10.1111/cogs.13226

Extracting Low-Dimensional Psychological Representations from Convolutional Neural Networks

Aditi Jha,[a,b] Joshua C. Peterson,[c] Thomas L. Griffiths[c,d]

[a]*Department of Electrical and Computer Engineering, Princeton University*
[b]*Princeton Neuroscience Institute, Princeton University*
[c]*Department of Computer Science, Princeton University*
[d]*Department of Psychology, Princeton University*

Received 24 May 2022; received in revised form 21 November 2022; accepted 23 November 2022

Abstract

Convolutional neural networks (CNNs) are increasingly widely used in psychology and neuroscience to predict how human minds and brains respond to visual images. Typically, CNNs represent these images using thousands of features that are learned through extensive training on image datasets. This raises a question: How many of these features are really needed to model human behavior? Here, we attempt to estimate the number of dimensions in CNN representations that are required to capture human psychological representations in two ways: (1) directly, using human similarity judgments and (2) indirectly, in the context of categorization. In both cases, we find that low-dimensional projections of CNN representations are sufficient to predict human behavior. We show that these low-dimensional representations can be easily interpreted, providing further insight into how people represent visual information. A series of control studies indicate that these findings are not due to the size of the dataset we used and may be due to a high level of redundancy in the features appearing in CNN representations.

Keywords: Similarity judgments; Categorization; Psychological representations; Neural networks; Deep learning; Interpretability

1. Introduction

Over the last decade, convolutional neural networks (CNNs) have been extremely successful in the field of computer vision, resulting in good performance on image classification and related tasks (He, Zhang, Ren, & Sun, 2016; Krizhevsky, Sutskever, & Hinton, 2012; Simonyan & Zisserman, 2015). The ability of CNNs to construct useful

Correspondence should be sent to Aditi Jha, PNI 238A, Princeton Neuroscience Institute, Washington Road, NJ 08540, USA. E-mail: aditijha@princeton.edu

representations of images from large image datasets has been of particular interest to cognitive scientists. CNN representations have been demonstrated to be highly predictive of multiple aspects of human visual perception (Kubilius, Bracci, & Op de Beeck, 2016; Lake, Zaremba, Fergus, & Gureckis, 2015). This utility has led to their representations being increasingly used to understand cognitive processes (Cichy & Kaiser, 2019; Kriegeskorte, 2015). For example, Sanders and Nosofsky (2020) show that CNNs can be trained to predict psychologically relevant dimensions of images identified through multidimensional scaling.

The representations of images used in these analyses are based on the activation of hidden units in extremely large artificial neural networks, typically resulting in thousands of potential features. This rich feature space is often thought to be the reason behind the success of CNNs in reproducing high-level cognitive phenomena. However, another possibility is that the large feature sets of CNNs contain psychological features— those learned by humans in the same task settings—as a subset. Recent work by Peterson, Abbott, and Griffiths (2018) and Battleday, Peterson, and Griffiths (2020) has shown that CNN representations are capable of encoding psychological information in the context of similarity and categorization tasks. A natural question that follows is whether this psychologically relevant information spans the entire set of CNN features or only a smaller subset. If psychological features are only a subset of CNN features, the remaining CNN features could be encoding redundant or irrelevant information. This would suggest that CNN representations may be compressible into lower dimensional representations while capturing psychologically relevant information. Typically, CNN representations are very high dimensional making it very hard to interpret them. However, compressing them into low-dimensional representations would enable us to easily interpret them and gain scientific insights into how humans represent visual stimuli.

Here, we examine how the dimensionality of representations derived from CNN features influences predictions of human behavior in two tasks. We first study similarity judgments, which are classically used in cognitive science to derive psychological representations. Peterson et al. (2018) showed correspondences between similarities in CNN representations and human similarity judgments for natural images, suggesting that CNN representations can be adapted to reflect human psychological representations. In the current work, we propose a method inspired by previous work by Rumelhart and Todd (1993) that allows us to find the low-dimensional projections of CNN representations that best capture human similarity judgments. Surprisingly, our method reveals that human similarity judgments can be predicted with two orders of magnitude fewer dimensions than those typically contained in CNN representations, indicating that psychological information is encoded by a small number of dimensions. To further study these low-dimensional psychological representations, we probe their individual dimensions, finding that most of them are interpretable. In particular, we show that broad visual categories are given more importance by our model than finer ones captured in subsequent dimensions, in line with the hierarchical structure often found to characterize human cognition (Cohen, 2000; Rogers & McClelland, 2004).

The second task we consider is categorization, another central cognitive function that has been studied for decades. The two classic modeling frameworks for human categorization are prototype models (Reed, 1972) and exemplar models (Medin & Schaffer, 1978; Nosofsky, 1984). Recent works (Battleday, Peterson, and Griffiths, 2020; Singh, Peterson, Battleday, &

Griffiths, 2020) have found that incorporating CNN representations into exemplar and prototype models can predict human categorization behavior for natural images. Building on their approach, we develop a framework for obtaining low-dimensional projections of CNN representations which also capture this behavior. Our analysis reveals that human categorization on the commonly used CIFAR-10 dataset can be predicted using representations with fewer than 10 dimensions.

To better understand the ability of low-dimensional representations derived from CNNs in predicting human similarity judgments and categorization, we conduct three control studies, each designed to control for a specific factor. First, we address the concern that our results may reflect the fact that the similarity datasets we used must necessarily focus on a relatively small number of images (in this case, each dataset contains 120 images). To rule out this possibility, we conduct a simulation in which we apply the same analysis to similarity scores derived directly from CNN representations themselves, showing that the inferred dimensionality of the resulting representations does not scale significantly with an increase in the number of images. Our analysis shows that low-dimensional representations are sufficient even in this case, suggesting that there is a high level of redundancy in CNN representations in general—that they are highly compressible. Our second control study follows up on this observation by evaluating the dimensionality of CNN representations required to perform standard image classification on the CIFAR-10 dataset. We find that fewer than 10 dimensions are sufficient to obtain comparable classification accuracy to full CNN representations. We then analyze the effect of class diversity and the number of basic-level categories on the dimensionality of CNN representations to ensure that low dimensionality is not merely a function of a number of categories. We find that fewer than 10 dimensions encode images sufficiently well to preserve classification accuracy (within 1%) on CIFAR-100 which has 100 classes, and 64 dimensions are sufficient to preserve classification information on ImageNet-100 which is more diverse at the level of basic categories. Finally, we show that this compressibility is not a function of our model but of the CNN representations. We observe that if we use word vectors as inputs to our model, we are not able to find low-dimensional representations of the inputs that are just as predictive of similarity judgments as the original input.

Overall, our results suggest that CNN representations contain a high level of redundancy, when using them to understand perceptual representations in humans. We show that we can find low-dimensional representations that can generate meaningful predictions about human behavior. Such low-dimensional representations overcome the interpretation barrier that typical CNN representations pose and can be probed to better understand the features that people use to solve visual cognitive tasks. Obtaining a succinct psychological representation of visual stimuli also makes several downstream analyses computationally less demanding, including correlating psychological representations with neural representations in the brain (Kriegeskorte, Mur, & Bandettini, 2008; Mur et al., 2013).

2. Convolutional neural networks

CNNs are a class of artificial neural networks tailored to process visual data. The goal of CNNs is to generate rich features from the input images, which can then be used to perform a range of downstream tasks. CNNs assume a specific structure in the input and are designed to exploit spatial correlations in images as well as to provide translation-invariant representations.

One defining feature of CNNs is the convolutional filter, which slides along the input image creating a new feature map of the image. This operation is mathematically equivalent to performing a convolution between the filter and the image. Typically, filters serve as feature detectors (such as edge detectors, blob detectors, or more complex object detectors), where the output feature map represents how strongly a feature is detected in a certain region of the image. Each convolutional layer contains multiple such filters, where each filter slides along the image providing its corresponding feature map. This is the first operation on an input image. Convolutional layers are typically followed by pooling layers, which perform a downsampling operation on the generated feature map. The most popular form of pooling is called max-pooling, where the feature map is downsampled such that the maximum unit in a specific region of the feature map (say of size 2×2) represents the entire region. An alternative to max-pooling is average pooling where small regions are represented by their means in the output feature map. The pooling filters are much smaller than the image itself, thus helping in generating translation invariant representations. The first few convolutional layers detect simple features (such as lines, texture), while the deeper layers detect more complex, high-level features (such as objects). After several stages of convolution and pooling, CNNs often employ fully connected layers at the end. Fully connected layers, where every node in a layer is connected to every other node in the next layer, contain a certain number of nodes, each with an activation value for the image. They are often followed by task-specific layers which use the image embedding for specific visual tasks, such as image classification. CNNs are trained to optimize performance on this visual task, and in the process they derive features that can be used to represent the input image in a diverse set of tasks.

Formally, for an input image x, the last fully connected layer's (also called the embedding layer) representation $F(x)$ contains a multidimensional feature representation of the image. This representation is rich, invariant to translations and can be used for downstream tasks. In all our experiments, we use this embedding layer representation for our images extracted from specific CNNs.

3. Analyzing similarity judgments

Similarity judgments have traditionally been used to study human psychological representations (Shepard, 1980). In a recent work, Peterson et al. (2018) showed correspondences between similarities in CNN representations and human similarity judgments for natural images. They found that while out-of-the-box CNN representations are only partially reflective of human psychological representations, they can be adapted to support a more fine-grained correspondence. The best performance was obtained with a representation that has 4,096 dimensions. In this section, we will determine how many of these dimensions are required to capture human similarity judgments.

Our approach is inspired by the classic work of Rumelhart and Todd (1993), who constructed a neural network to predict human similarity judgments. The model takes vectors representing two stimuli as input and outputs a similarity judgment. The hidden layer is of a lower dimensionality than the input, resulting in a compressed representation. Extending this idea to modern CNNs, our method—which we call Similarity-driven Dimensionality Reduction (SimDR)—reduces the CNN representations of images to a low-dimensional space which is optimal for predicting human similarity judgments.

3.1. Methods

3.1.1. Stimuli

Peterson et al. (2018) collected six human similarity datasets for natural images drawn from the following domains: animals, vehicles, vegetables, fruits, furniture, and a dataset encompassing a variety of domains ("various"). Participants rated each pair of images on a scale of 0–10 (where higher numbers imply greater similarity). For each domain, there are approximately 10 human similarity ratings for each pair of images. With a total of 120 images per domain, this results in approximately 71,400 similarity judgments per domain ($(\binom{120}{2}$pairs $\times$ 10 trials per pair).

3.1.2. Similarity-driven dimensionality reduction

Peterson et al. (2018) showed that the final, fully connected representation layer of VGG-19 (Simonyan & Zisserman, 2015) is most predictive of human similarity judgments; hence, we use the same 4096-dimensional (4096-d) VGG-19 representations for all our experiments. The task of obtaining low-dimensional representations of images which capture factors underlying human similarity judgments is thus a matter of projecting VGG-19 representations to a low-dimensional space that still predicts human similarity judgments.

We solve this problem by constructing a simple linear neural network. We use a linear network since we only want to extract the useful dimensions out of CNN representations, not to modify them in any way. For two input images, x_i and x_j, we first obtain their corresponding 4096-d VGG-19 representations, $F(x_i)$ and $F(x_j)$. For compactness, we will refer to them as F_i and F_j, respectively. They are passed through a single linear layer, B, of small width (i.e., a bottleneck layer) which projects them to a lower dimensional space, with $B(F) = W_B F$, where W_B are the weights of the bottleneck layer. The weights of this bottleneck layer are shared, such that we obtain the vectors $B(F_i)$ and $B(F_j)$, respectively, for the two images. The inner product of the outputs of the bottleneck layer is then used to predict similarity rating for the input pair (Fig. 1), $\widehat{s_{i,j}} = B(F_i)^\top B(F_j)$. The inner product is our representational similarity measure, which contrasts with Rumelhart and Todd (1993), and more directly generalizes the method of Peterson et al. (2018). We refer to the mean human similarity judgment between the given pair of images as $s_{i,j}$. The weights of the bottleneck layer are learned by back-propagating the loss incurred during the prediction of human similarity judgments, hence optimizing the projected representations to predict human similarity judgments (the CNN representations of the images F_i and F_j are held fixed). This

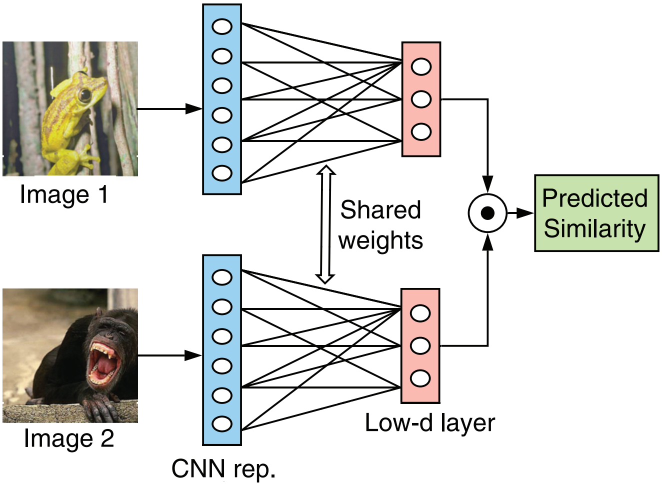


Fig. 1. Overview of SimDR. CNN representations for an image pair are down-projected using a shared low-dimensional bottleneck layer. An inner product of the outputs gives predicted similarity rating for the input pair.

results in the following loss function for a given image pair (i, j):

$$L_{i,j} = ||\widehat{s}_{i,j} - s_{i,j}||^2 = ||B(F_i)^\top B(F_j) - s_{i,j}||^2, \tag{1}$$

which is optimized with respect to the weights in the bottleneck layer B.

The loss function in Eq. 1 is equivalent to $||F_i^\top W_B^\top W_B F_j - s_{i,j}||^2$. In other words, we are learning a low-rank matrix $(W_B^\top W_B)$ to map the feature vectors to similarity judgments. This contrasts with the method of Peterson et al. (2018), which learns a weight for each of the 4,096 input dimensions (i.e., a diagonal matrix in place of any general $W_B^\top W_B$). Gershman and Tenenbaum (2015) used a similar approach to learn phrase similarity, using a full matrix W in place of $W_B^\top W_B$, without a low-rank constraint. Our low-rank formulation allows us to specify the dimensionality of the weight matrix W_B and learn a low-dimensional representation for images, which is our core objective. Our approach is also distinct from principal component analysis (PCA; Pearson, 1901), which focuses on capturing variation in the inputs rather than just the information that is relevant to human similarity judgments (and thus may inflate the estimated dimensionality).

We trained a separate model for each dataset. CNN feature vectors were normalized such that their norms were one. We used mean squared error as the loss function with L2 regularization on the weights of the bottleneck layer, W_B to train each model:

$$L = \frac{1}{2} \sum_i \sum_j ||B(F_i)^\top B(F_j) - s_{i,j}||^2 + \lambda ||W_B||^2. \tag{2}$$

L2 regularization adds the additional term $\lambda||W_B||^2$, which drives the weights to small values such that the overall similarity prediction function is smooth and consequently avoids overfitting. We selected the L2 coefficient, λ, between 10^{-3} and 10^3 by six-fold cross-validation.

Table 1
R^2 scores for all datasets (SimDR and PCA values are for the bottleneck layer of size 64, the results from Peterson et al. (2018) reflect the CV (cross-validation) control values from their Table 1)

Dataset	Raw	Peterson et al. (2018)	SimDR	PCA
Animals	0.58	0.74	0.64	0.47
Vehicles	0.51	0.58	0.57	0.51
Fruits	0.27	0.36	0.30	0.27
Furniture	0.19	0.35	0.33	0.28
Various	0.37	0.54	0.50	0.31
Vegetables	0.27	0.34	0.30	0.32

Abbreviations: PCA, principal component analysis; SimDR, Similarity-driven Dimensionality Reduction.

We measured the accuracy of the similarity predictions as compared to human similarity ratings using the coefficient of determination, R^2. Finally, we varied the number of nodes in the bottleneck layer, B, in the range of 1–64. For each setting of the bottleneck layer's width, we report the average validation R^2 score using six-fold cross-validation over the full set of 120 images (and at the best λ selected by a separate run of six-fold cross-validation over the same set of images due to the small size of our datasets). We also compare this approach with a simple unsupervised baseline that obtains low-dimensional representations by running PCA over the input VGG-19 representations. These low-dimensional representations are then transformed using the method of Peterson et al. (2018) to best predict similarity ratings. As above, we vary the number of principal components in the range of 1–64.

3.2. Few dimensions predict similarity judgments

For all datasets, we observe that the proportion of variance in similarity scores accounted for by the model (R^2) obtained using SimDR score at 64 dimensions is higher than that of the raw (untransformed) CNN representations (Table 1). The PCA-based model performed worse than SimDR for all datasets (except for the *vegetables* dataset), suggesting that supervision is much more selective of the human-relevant dimensions. We also observe that the prediction performance of SimDR quickly saturates as the number of dimensions increases beyond 10–20, approaching the prediction performance obtained using all VGG-19 features (Fig. 2; dashed lines). Notably, the *animals* dataset requires only six dimensions to achieve an R^2 score of 0.6 while the *various* dataset achieves an R^2 of 0.49 at six dimensions. These results strongly suggest that human similarity judgments can be captured by considerably fewer dimensions (by at least two orders of magnitude) than those comprising VGG-19 representations, and more generally that psychological representations as measured by similarity experiments are much lower dimensional than CNN representations. Additional evidence for this can be seen in the intrinsic dimensionality of the CNN representations themselves without respect to human judgments. Fig. 3 illustrates this using PCA: cumulative variance explained is shown as a function of the number of components for each dataset. Notably, the dimensionality elbow is both longer and later than those in Fig. 2. Interestingly, CNNs also appear to assign equal dimensionality to all datasets (except *vehicles*), apparently much unlike humans (Fig. 2).

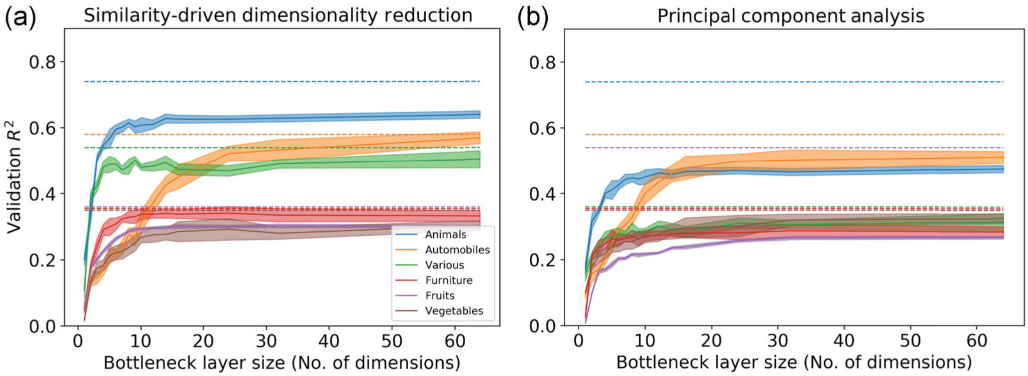


Fig. 2. Explained variance (R^2) of our models in predicting human similarity judgments on each dataset. The dashed lines correspond to the prediction performance in Peterson et al. (2018) when all input dimensions are used. (A) shows the R^2 score (averaged over 10 runs, error bars reflect the 95% confidence interval over the 10 runs) for our end-to-end similarity prediction model, SimDR. (B) shows the R^2 score (also averaged over 10 runs, with a 95% confidence interval) when we used PCA to obtain low-dimensional representations of images from their corresponding VGG-19 representations, and measured similarity between these low-dimensional representations. In both cases, for each run, we report the mean validation R^2 score using six-fold cross-validation.

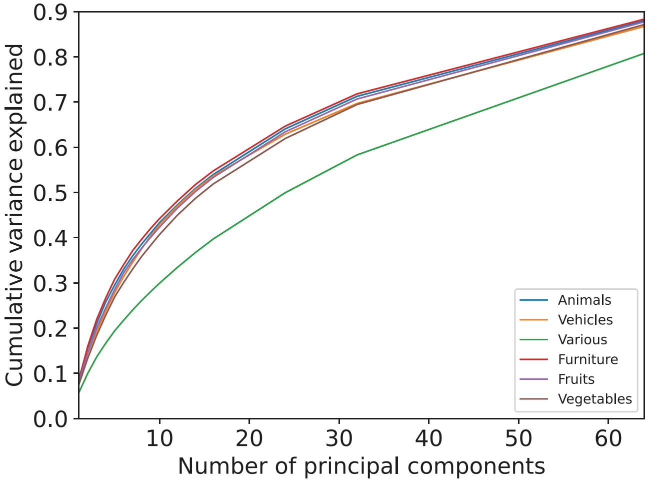


Fig. 3. Cumulative variance explained in the VGG-19 representations as a function of the number of principal components used to reconstruct them, for each dataset. Unlike Fig 2, the amount of variance captured in these representations does not saturate at a small dimensionality. Hence, the inherent variance in VGG-19 representations is captured by a large number of dimensions.

3.3. *Interpretation of low-dimensional features*

Now that we have demonstrated that low-dimensional representations are sufficient to predict similarity judgments, we can attempt to interpret the reduced set of dimensions. For this experiment, we focus on the top three datasets based on an R^2 score—*animals*, *vehicles*, and

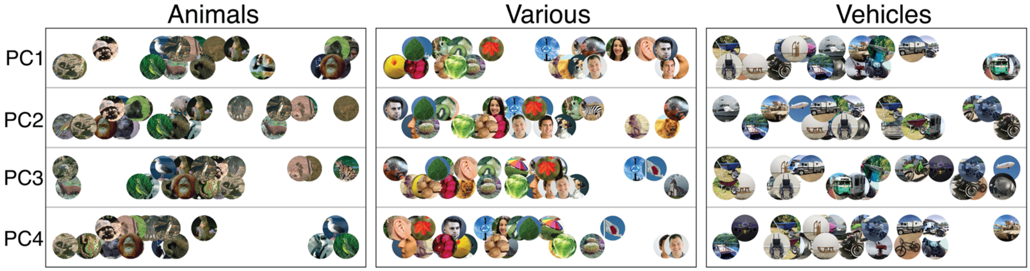


Fig. 4. Image embeddings visualized along the top four principal components of low-dimensional SimDR representations. Several of these dimensions capture meaningful concepts. Most notably, the first PC of the *animals* dataset distinguishes between mammals and non-mammals, and the first PC of the *various* dataset separates inanimate objects from dogs and humans.

various. As mentioned above, SimDR achieves an R^2 score of 0.6 on the *animals* dataset using six dimensions, an R^2 score of 0.49 on the *various* dataset using six dimensions, and an R^2 score of 0.45 on *vehicles* dataset using 16 dimensions. We fix these as the bottleneck layer sizes for each of these datasets. The aforementioned dimensions for each of the three datasets are chosen by visually identifying an elbow in performance (Fig. 2) such that the rate of increase in an R^2 score is small beyond this point. We want to understand these individual dimensions; however, they may not be orthogonal. To address this, we further orthogonalize our low-dimensional representations using PCA to ensure that each dimension encodes unique information. We then take the top few dimensions which explain most of the variance for each dataset. This contrasts with the use of PCA above to produce a baseline-reduced feature set in that it is performed after supervised dimensionality reduction and is only for the sake of rotating the resulting set of factors.

For each of the three datasets, we visualize image embeddings along the top four principal dimensions of the low-dimensional features learned via SimDR (Fig. 4). We visualize validation images for a single fold (out of the six cross-validation folds), though we observe that the dimensions were consistent across all folds in terms of capturing the same concepts (see the Appendix). We observe that the first dimension for each dataset appears to be largely continuous and captures broad categories. In the *animals* dataset, this dimension goes from non-mammals to mammals. The first dimension of the *various* dataset goes from inanimate objects to dogs and humans. Finally, the first dimension of the *vehicles* dataset shows a gradation from vehicles with two or no wheels (e.g., sled, wheelchair) to those with four wheels (e.g., trucks, buses), though the interpretation in this case is not as evident, which may stem from the low variance (12%) captured by the top component. Some of the other principal components are also apparently interpretable and interesting. For example, the second principal component of the *vehicles* dataset distinguishes water transport from land transport, the third principal component of the *various* dataset distinguishes natural things from artificial ones, while the fourth dimension in the *animals* dataset distinguishes birds from non-birds. Each of these individual dimensions captures a different taxonomic relationship, suggesting that such relationships are important factors in determining similarity judgments of natural images.

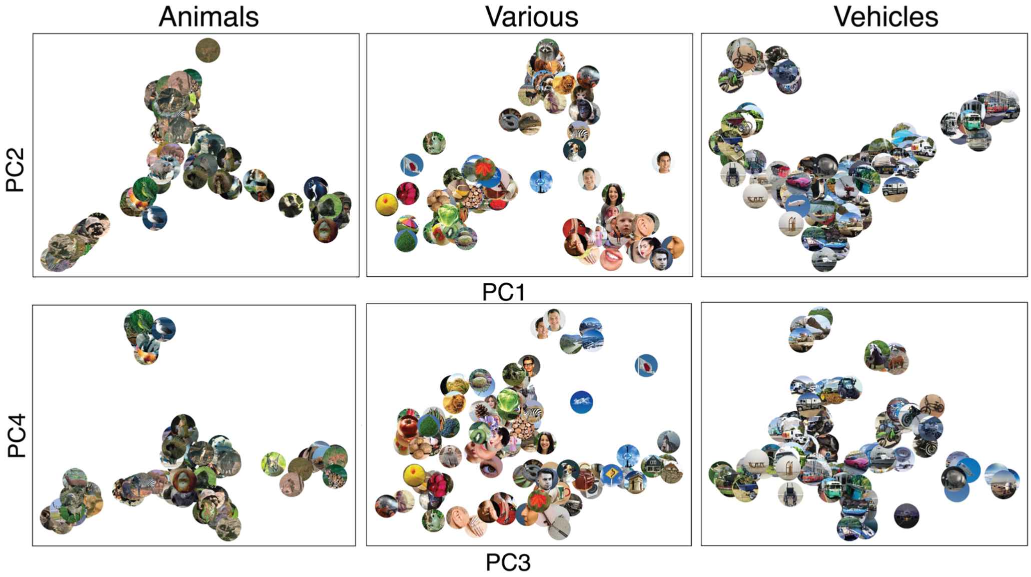


Fig. 5. Examples of image embeddings for three datasets using the top principal components (PCs) of the SimDR representations. Here, the image embeddings are visualized along two PCs (PC1 vs. PC2 and PC3 vs. PC4). These plots reflect meaningful clusters formed by projecting images along a pair of PCs. Some examples are birds in the *animals* dataset (see PC3 vs. PC4, top), animals in the *various* dataset (see PC1 vs. PC2, top center), and *airplanes* in the vehicles dataset (see PC3 vs. PC4, bottom right).

As an alternative visualization strategy, we explore two-dimensional projections of the image representations along two of the top four principal components in Fig. 5. These plots are useful in observing clusters of images formed by a combination of principal components, where each cluster tells us what kinds of images are considered similar by the model. Echoing (Peterson et al., 2018), we observe clusters for herpetiles (reptiles and amphibians), primates, birds, wild cats, rodents, and grazers in the *animals* dataset.

We see clusters for human faces and body parts, animals, vegetables, houses, and natural things in the *various* dataset. The *vehicles* dataset shows distinct clusters for trains, bikes, horses, airplanes, and tanks.

3.4. Hierarchical similarity and bottleneck effects

Next, we analyze the effect of changing the width of the bottleneck layer. We know that increasing the width improves prediction performance. Here, we are interested in interpreting the information captured by different bottleneck sizes.

To visualize this, we explore hierarchical clustering (Shepard, 1980) of judgments from the *animals* dataset. Fig. 6 shows via dendrograms that when the bottleneck layer in SimDR has two dimensions, two clusters—primates and non-primates—are formed. This suggests that belonging to the primate group is the most important trait influencing similarity judgments in the *animals* dataset, which is encoded in as little as two dimensions. At a bottleneck size of 6,

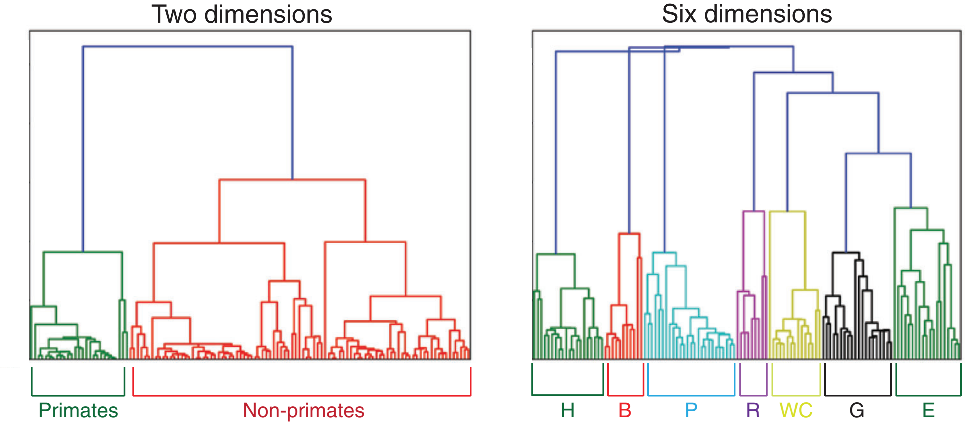


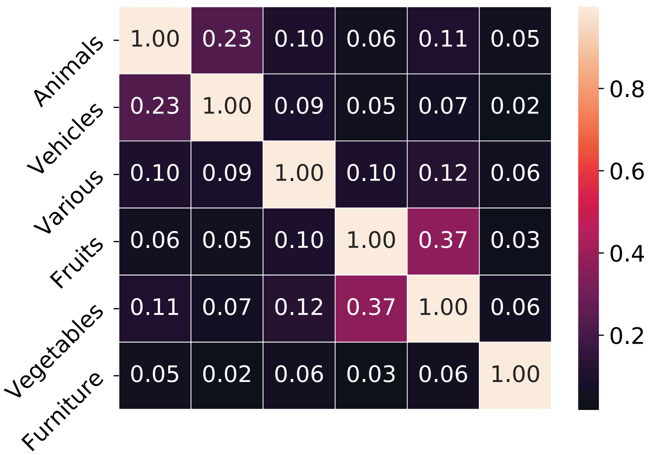
Fig. 6. Dendrograms produced by hierarchical clustering for two-dimensional representations and six-dimensional representations on the *animals* dataset. The representations were obtained from SimDR. We find a hierarchical organization for the image representations, where two–dimensional representations capture high-level categories while the six-dimensional representations capture finer distinctions within these categories. *H: Herps, B: Birds, P: Primates, R: Rodents, WC: Wild cats, G: Grazers, E: Dogs, Bears and Large animals*.

however, a further hierarchical structure can be seen where many more categories are present. At intermediate sizes between 2 and 6, additional clusters continue to emerge (not shown). The hierarchical structure formed by the six-dimensional representations is closely related to that formed using human similarity data in Peterson et al. (2018).

We further observe that increasing the bottleneck width introduces additional categorical distinctions in other datasets too. For the *various* dataset, at a bottleneck width of four, we observe distinct clusters for animals and humans (and their body parts). In the case of the *vehicles* dataset, four-dimensional bottleneck layer representations preserve distinctions based on wheels. Hence, these are primary traits influencing similarity judgments which are captured at small bottleneck widths. These results motivate a hierarchical organization of factors underlying human similarity judgments in our model, providing empirical results consistent with mathematical theories of hierarchical semantic knowledge organization in neural networks (Saxe, McClelland, & Ganguli, 2019).

3.5. Shared features across domains

We have seen that each of the six individual SimDR models can discover low-dimensional representations which are predictive of similarity judgments separately for each domain. A natural question that follows from this is whether the dimensions learned by these models trained on specific domains are also shared across domains. Translating this into the framework of human judgments, the question we pose is the following: Do different domains share factors underlying human similarity judgments?



Animals	1.00	0.23	0.10	0.06	0.11	0.05
Vehicles	0.23	1.00	0.09	0.05	0.07	0.02
Various	0.10	0.09	1.00	0.10	0.12	0.06
Fruits	0.06	0.05	0.10	1.00	0.37	0.03
Vegetables	0.11	0.07	0.12	0.37	1.00	0.06
Furniture	0.05	0.02	0.06	0.03	0.06	1.00

Fig. 7. Interdomain relatedness (R^2) as measured by regularized CCA, between all pairs of image domains. Higher R^2 between a pair of domains suggests a higher degree of shared information between the domains, as captured by representations obtained from SimDR. In particular, the representations of *fruits* and *vegetables* share information, the same is also true for the *animals* and the *vehicles* datasets.

3.5.1. Canonical correlation analysis

We use L2-regularized canonical correlation analysis (CCA; Bilenko & Gallant, 2016) to evaluate the degree of shared information or factors between low-dimensional representations belonging to any two domains. From each of the six models trained on individual domains, we obtain 64-dimensional representations for all pairs of images (from all six datasets). We then perform regularized CCA on 64-dimensional representations from every pair of domains.

We observe in Fig. 7 that the R^2 score is highest for *fruits* and *vegetables*, followed by *animals* and *vehicles*. This implies that the representations formed for fruits and vegetables overlap. The same is true for *animals* and *vehicles* datasets. While it seems reasonable for *fruits* and *vegetables* to share common factors for similarity, the relationship between *animals* and *vehicles* is less clear, although we suspect it may have something to do with common backgrounds (which often contain scene information such as grass, sky, and water, unlike our other categories).

3.5.2. Domain-agnostic SimDR

To determine whether a more general set of dimensions could be learned that generalizes across domains, we trained a SimDR model on image pairs from all six datasets using six-fold cross-validation. We compared this to models trained on individual domains and tested on all others to assess how they generalize on their own. The results, shown in Fig. 8 (panel A), reveal that the pooled model nears saturation at a few hidden dimensions. Hence, even with a diverse dataset, few dimensions are enough to predict similarity judgments. Next, we see that the domain-specific models do not generalize well when tested on all datasets, lending credibility to our earlier claim that these models learn dimensions which are specific to individual domains. Lastly, Fig. 8 (panel B) shows the performance of the pooled model in

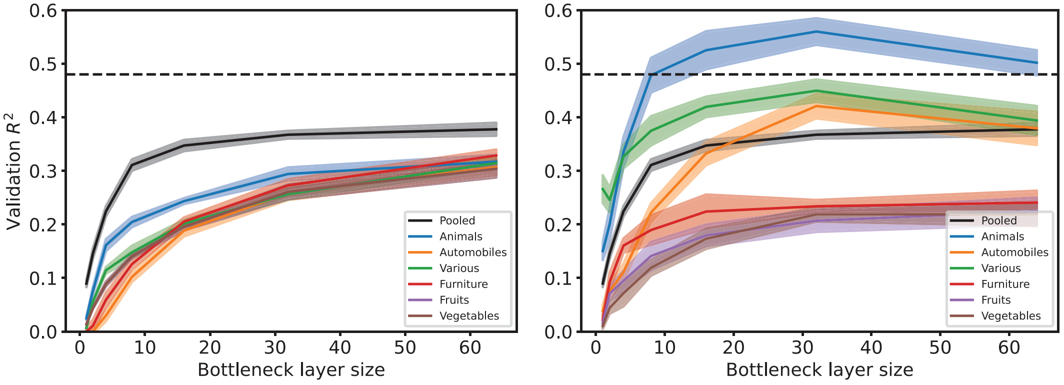


Fig. 8. Evaluation of SimDR on a pooled set of images from all of the six datasets. (A) Performance of models tested on all domains (with varying bottleneck layer size). (B) Performance of pooled model tested on individual domains and on all domains (with varying bottleneck layer size). The dashed line shows the performance of the model trained on all domains in Peterson et al. (2018). Solid lines correspond to the pooled model tested on different datasets; they are averaged over 10 runs and the error bars reflect 95% confidence intervals over the 10 runs.

predicting individual domains and reveals that certain domains (*animals*, *vehicles*, *various*) are well-explained by general features learned from the pool of all domains, while others require more domain-specific features (*vegetables*, *fruits*, *furniture*).

These results show that CNN representations can be transformed to lower dimensional representations—where interpretation is far less cumbersome—while still being predictive of human similarity judgments. We observe that only a few dimensions are required to predict psychological representations as opposed to the thousands of dimensions for a full set of CNN features. This suggests that CNN representations contain only a subset of features that are directly relevant to capturing psychological information, and many of the features discovered by CNNs may be redundant.

4. Analyzing human categorization

Categorization is a central and extensively studied cognitive task, with an extensive literature providing models of categorization evaluated with simple visual stimuli (Medin & Schaffer, 1978; Nosofsky, 1984; Reed, 1972). In a recent work, Battleday et al. (2020) explored categorization involving complex natural images by combining CNN representations with existing cognitive models. They used the CNN representations of natural images as input to classic categorization models (prototype and exemplar models) and were able to predict human categorization behavior. In this section, we complement the analysis of similarity data presented above by developing a pipeline to reduce the dimensionality of CNN representations while training them simultaneously to predict human categorization.

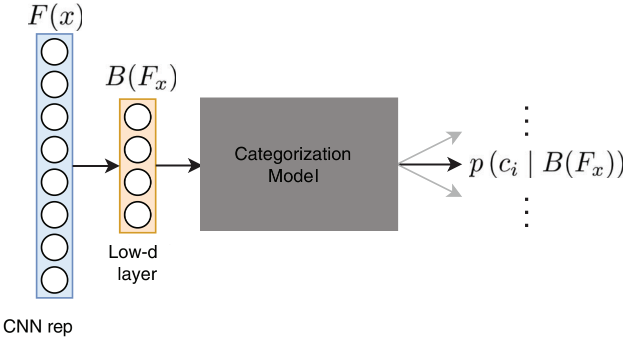


Fig. 9. Overview of CatDR. CNN representation, $F(x)$ (shortened to F_x), for an image, x is down-projected using a low-dimensional bottleneck layer, $B(F_x)$. This serves as input to a categorization model (prototype or exemplar), which outputs the posterior probabilities of classes.

4.1. Methods

4.1.1. Stimuli

To train our categorization models, we used the CIFAR-10H behavioral dataset developed by Peterson, Battleday, Griffiths, and Russakovsky (2019) and used by Battleday et al. (2020). This dataset contains 511,400 human categorization decisions made over 10,000 natural images. The images were taken from the test set of CIFAR-10 which encompass 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Participants were asked to categorize every image as accurately as possible. Approximately 50 judgments per image were collected, resulting in a probability distribution over the 10 classes for each image.

4.1.2. Categorization-driven dimensionality reduction

We extracted CNN representations for the 10,000 images from a ResNet-18 (He et al. ,2016) model trained on CIFAR-10. We used ResNet-18 because its parameter space is just constrained enough to allow easy training but also provides high accuracy on the CIFAR-10 dataset. Moreover, related work by Singh et al. (2020) demonstrates the ability of ResNet-18 to predict human categorization decisions. After obtaining CNN representations corresponding to our stimuli, the task of extracting low-dimensional representations that are predictive of human categorization can be carried out in two steps: a linear low-dimensional projection of the CNN representation followed by categorization using prototype and exemplar models. Hence, the categorization models receive a low-dimensional projection of the CNN representations as input, which we also refer to as the bottleneck layer. This entire framework can be optimized together such that the bottleneck layer captures dimensions that are predictive of human categorization decisions. Fig. 9 shows an illustration of the model, which we refer to as Categorization-driven Dimensionality Reduction (CatDR). The CNN representation, $F(x)$, of an image x, is passed through a fully connected bottleneck layer resulting in a low-dimensional representation, $B(F(x))$ (shortened to $B(F_x)$). This is a linear projection, such that $B(F_x) = W_B F_x$, where W_B are the weights of the bottleneck layer. This bottleneck representation serves as input to a categorization model which outputs per-class probability given the bottleneck representation, $p(c_i \mid B(F_x))$. CatDR is trained simultaneously to learn the parameters of the categorization model as well as down-projection weights for the bottleneck layer.

Prototype models Categorization models (Anderson, 1991; Ashby & Alfonso-Reese, 1995) aim to compute the probability that a stimulus x belongs to a class $c_i \in C$, where C represents the set of classes. For a prototype model (Reed, 1972), this is computed by comparing the stimuli x to the prototypes or summary representations for each class μ_i. This similarity measure $S(x, \mu_i)$ is usually taken to be an exponentially decreasing function of the distance between the stimulus and the prototype such that

$$S(x, \mu_i) = e^{-d(x,\ \mu_i)}. \tag{3}$$

Here, $d(x,\ \mu_i)$ represents the distance between x and μ_i, which can simply be the squared Euclidean distance. More generally, one can think of each class as following a Gaussian distribution with mean μ_i and covariance Σ_i, resulting in the more general expression

$$S(x, \mu_i) = e^{-d(x;\ \mu_i, \Sigma_i)}, \tag{4}$$

where $d(x;\ \mu_i, \Sigma_i) = (x - \mu_i)\Sigma_i^{-1}(x - \mu_i)^\top$ or the Mahanabolis distance between the stimuli and the class mean, given the class covariance. Similarity can then be related to the probability distribution over classes, $p(c_i \mid x)$, using the Luce–Shepard choice rule (Luce, 1959)

$$p(c_i \mid x) = \frac{S(x, \mu_i)^\gamma}{\sum_{j \in C} S(x, \mu_j)^\gamma}, \tag{5}$$

where γ is a learnable parameter.

Applying CatDR in this model, we start with CNN representations of the image and then try to learn the low-dimensional or bottleneck layer. The probability distribution over classes is now given by $p(c_i \mid B(F_x))$, where $B(F_x) \in \mathbb{R}^d$ is the bottleneck representation of the stimulus x. For a prototype CatDR model, we have

$$p(c_i \mid B(F_x) \propto S(B(F_x), \mu_i)^\gamma. \tag{6}$$

Also, the class mean μ_i and Σ_i are now low-dimensional, $\mu_i \in \mathbb{R}^d$, and $\Sigma_i \in \mathbb{R}^{d \times d}$. Our objective is to learn the bottleneck weights W_B and the categorization parameters: γ, μ_i, and Σ_i simultaneously. A fully expressive covariance Σ_i is hard to learn, and so often categorization models make assumptions about Σ_i. In this paper, we test linear and quadratic prototype models. Linear prototype models assume that all classes share the same covariance matrix and fix Σ_i to be a diagonal, which means that each feature varies independently. Quadratic prototype models also assume Σ_i to be diagonal, but it is allowed to vary over classes. Further, the means μ_i are initialized as random projections of the training class means to $\mathbb{R}^d$.

Exemplar models In classic exemplar models (Nosofsky, 1984), the probabilities of choosing each category are computed by comparing the stimulus, x, with all stored examples,

x' for a given category. This results in

$$p(c_i \mid x) \propto \Big(\sum_{x' \in c_i} S(x, x')\Big)^{\gamma}, \tag{7}$$

where $S(x, x') = e^{-\beta d(x, x')}$. We fix $d(x,\ x')$ to represent the squared Euclidean distance between the stimulus x and an exemplar x'. β is called the specificity parameter and is learned during training. The exemplar model does not restrict the class-structure to be Gaussian but instead allows for arbitrarily shaped class distributions by approximating them using multiple samples. In our low-dimensional approach, the inputs to the exemplar model are the bottleneck representations of the stimulus. Hence, we can modify the choice probabilities as follows:

$$p(c_i \mid B(F_x)) \propto \Big(\sum_{x' \in c_i} S(B(F_x), B(F_{x'}))\Big)^{\gamma}, \tag{8}$$

where again $B(F_x) \in \mathbb{R}^d$ is the output of the bottleneck layer with dimension d. Here, the exemplars are also passed through the bottleneck layer, such that now they are low dimensional. We set x' to be the training examples from the given category c_i. Hence, the exemplars that the categorization model works with are low-dimensional projections of these training samples x'. In this model, the objective is to learn the parameters γ, β, and the bottleneck layer weights W_B simultaneously.

4.1.3. Training and evaluation details

To train the models, we set the size of the bottleneck layer $\in [1, 128]$. We use the CIFAR-10H dataset and minimize cross-entropy loss using the Adam optimizer (Kingma & Ba, 2015) with a learning rate of $1e-4$ and a batch size of 256 for gradient updates. We perform five-fold cross-validation and report the averaged validation loss as well as the averaged validation classification accuracy. To compute classification accuracy, the category with maximum probability is selected as the predicted category.

4.2. Few dimensions are sufficient for predicting human categorization

We trained and evaluated prototype and exemplar versions of the CatDR model on CIFAR-10H using ResNet-18 representations which are in $\mathbb{R}^{512}$. We varied the size of the bottleneck layer in the range of 1–64 and observed that the validation loss as well as the classification accuracy start saturating after just eight dimensions in case of both prototype and exemplar models (Fig. 10). This means that only eight dimensions are sufficient for human categorization in the case of CIFAR-10 images. We also compared our results to the deep categorization models developed by Battleday et al. (2020), that is, without any bottleneck layer. We observe that with much fewer dimensions, our low-dimensional categorization models achieve a similar loss and accuracy. This is an important result, as it opens doors for interpretation of each of these dimensions in the future to infer the underlying criterion used by people to categorize images.

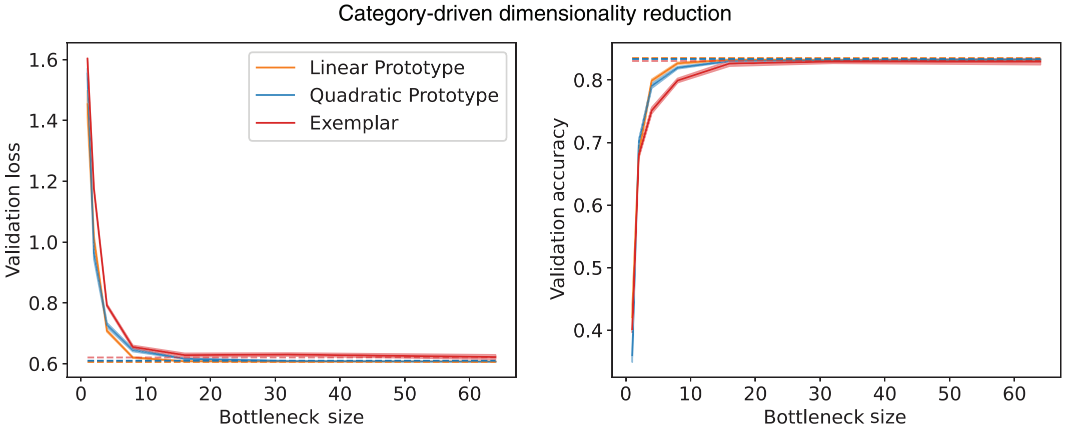


Fig. 10. Cross-entropy loss and classification accuracy (computed by selecting the highest probability class) as a function of the size of the bottleneck layer for variants of CatDRs. The dashed lines correspond to performance of the categorization model in Battleday et al. (2020), with all input dimensions and no bottleneck layer. Solid lines are averaged over 10 runs, with a narrow error bar reflecting the 95% confidence interval over these runs.

We also observe that in alignment with Battleday et al. (2020) prototype models tend to perform better than exemplar models. As they note, this can be a consequence of the fact that the CNN representations are designed specifically to be classified using linear boundaries which puts prototype models at an advantage as some are restricted to linear boundaries, unlike exemplar models. We also observe that low-dimensional quadratic prototype models, despite being more expressive, do not show any improvement over low-dimensional linear prototype models. This could, in part, be due to overfitting of the model parameters. Since the categorization models are being trained on far-lower dimensional inputs than if trained on raw CNN representations themselves, quadratic prototype models might be over-parameterized resulting in overfitting while training. Overall, the key takeaway here is the ability of low-dimensional CNN projections to predict human categorization behavior.

5. Control studies

Our results so far suggest that low-dimensional projections of CNN representations are sufficient to predict human similarity judgments and categorization behavior. We performed three control studies to further validate these findings. These studies suggest that CNN representations are inherently compressible, allowing us to identify low-dimensional projections which capture human behavior.

5.1. *Similarity prediction on large datasets*

Our experiments using SimDR are based on small datasets (120 images each), and so we wanted to rule out the possibility that our results could be due to dataset size. We performed a sanity check to do so. We randomly sampled images for each of the six categories used in

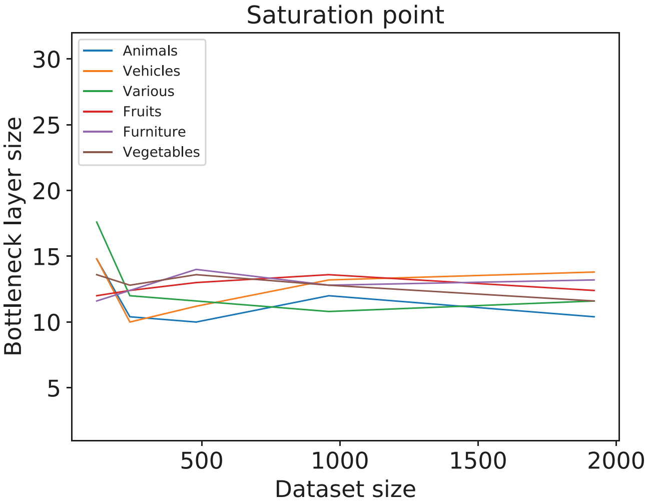


Fig. 11. Number of dimensions in the bottleneck layer at which the explained variance of our model in predicting similarity between CNN representations saturates.

our previous analysis (animals, vehicles, vegetables, fruits, furniture, and "various") from the larger Imagenet dataset (Deng et al., 2009) to create datasets with sizes ranging from 120 to 1920 images for each category. We then used similarity scores based on CNN representations of these images as a proxy for human similarity ratings to examine how dataset size affects inferred dimensionality. Specifically, we obtained pairwise similarity ratings for images in these datasets by computing cosine similarity between their corresponding VGG-19 representations. We then trained SimDR at varying bottleneck layer widths (1–64) to predict the pairwise similarity ratings for each of the six datasets. To visualize the results, we plotted the bottleneck layer size at which the validation R^2 score saturates (which we refer to as the saturation point) as a function of dataset size in Fig. 11. We observe that the saturation point for all datasets lies between 10 and 20 across varying dataset sizes. The fact that the saturation point does not significantly increase as a function of dataset size indicates that the low-dimensional solution that we found when predicting human similarity judgments is not due to the dataset size.

5.2. *Dimensionality constrained image classification at varying number of classes*

Our results on categorization motivated us to further investigate whether low-dimensional representations are unique to capturing human categorization, or if similarly low-dimensional representations can be used in the image classification tasks that CNNs were originally designed to solve. If this is the case, it would imply that some of the dimensions in CNN representations are redundant even for classification purposes.

To this end, we added a bottleneck layer between the final representation layer and softmax layer of ResNet-18. We varied the width of this bottleneck layer between 8 and 64 dimensions and trained these models from scratch on CIFAR-10's training dataset with 50,000 images

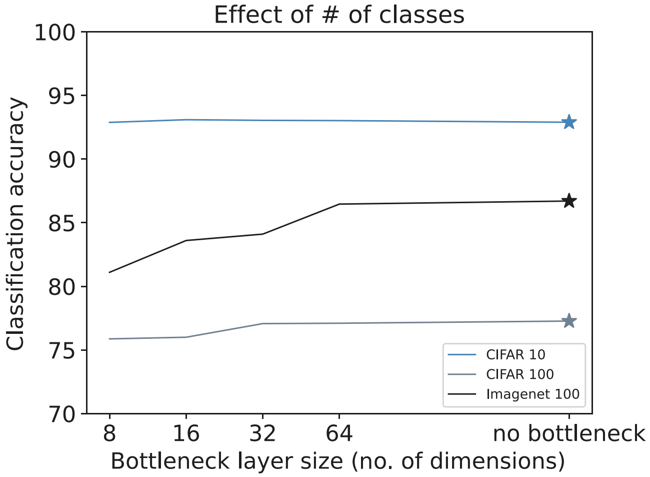


Fig. 12. Classification accuracy of ResNet-18 models with an additional bottleneck layer (added b/w the second last and the last layer) as a function of the width of this layer, on CIFAR-10 (blue), CIFAR-100 (gray) and ImageNet-100 (black). The stars correspond to the performance of ResNet-18 without any bottleneck (the layer before softmax in this case is 512-dimensional) on the two datasets.

and 10 classes. Surprisingly, we observed that the classification accuracy did not suffer (Fig. 12, blue trace) when compared to ResNet-18 without a bottleneck layer which has a 512-dimensional layer before the last classification layer (Fig. 12, blue star). This implies that even eight-dimensional representations are sufficient to encode classification information for CIFAR-10 images.

Next, we also wanted to analyze if the number of classes in a dataset has an effect on the number of dimensions needed to sufficiently encode the corresponding images. To do this, we repeated the above experiment for the CIFAR-100 dataset. CIFAR-100 has the same set of images as CIFAR-10, but with finer image categories resulting in 100 classes. We again trained our modified ResNet-18 model with varying bottleneck widths (between 8 and 64) on 50,000 images and tested them on the remaining 10,000 images. We compared the classification accuracy of these models with that of vanilla ResNet-18 trained and tested on the same set of images. We find that the classification accuracy drops only by 1% when using a bottleneck width of 8 (Fig. 12, gray trace) as compared to the model without any bottleneck (Fig. 12, gray star). At higher bottleneck sizes of 32 and, beyond, we see no drop in performance. This suggests that even when using datasets with much higher class diversity, we can obtain succinct low-dimensional representations of visual stimuli that are equally good for classification as the images themselves.

Furthermore, we also tested whether the number of basic-level categories in a dataset influences the number of dimensions needed to encode the images. We repeated the above experiments on ImageNet-100 (Chun-Hsiao Yeh, 2022), which is a subset of ImageNet-1k (Russakovsky et al., 2015), containing 100 random classes with 1300 training images and 50 test images per class. We find that the test classification accuracy of ResNet-18 saturates at a

Table 2
Cross-entropy loss of categorization models using 8t-dimensional representations extracted from modified ResNet-18 (with a bottleneck) and, of CatDR with an 8-dimensional bottleneck layer when trained on 512-dimensional ResNet-18 representations

Model Type	With 8-Dimensional Input	With 8-Dimensional Bottleneck Layer
Linear prototype	0.63	0.62
Quadratic prototype	0.65	0.64
Exemplar	0.71	0.65

Note. Results are averaged over 10 runs.

bottleneck of 64 dimensions and is comparable to what we obtain using the full embedding space (512 dimensions). At smaller bottleneck sizes, we do observe a drop in performance but the slope is very small. For example, at eight dimensions, we obtain an accuracy of 81.1% as compared to 86.7% at 512 dimensions. This indicates that while including more basic-level categories does require a higher dimensional representation to encode all classification-related information, performance is still high using < 10 dimensions and levels off at only 64-dimensional representations.

Finally, to examine whether the inferred low-dimensional representations can be used to predict human categorization behavior, we also used the eight-dimensional CNN representations extracted from the modified ResNet-18 as inputs to our categorization models (without any bottleneck layer). We found that the prototype models perform comparably to the models where the bottleneck layer was specifically trained to be predictive of human categorization. Hence, representations learned by the modified ResNet-18 for classification are also capable of predicting categorization decisions using prototype models (see Table 2). However, for exemplar models, the eight-dimensional representations fall short of being as predictive of human categorization decisions as CatDR trained on full CNN representations. This suggests that the exemplar models are able to capitalize on more customized bottleneck representations, as opposed to representations learned for image classification.

5.3. Predicting word similarity ratings

Our results suggest that CNN representations may contain a high level of redundancy in their inferred features. To evaluate whether this property is specific to these models, we conducted another set of experiments using models trained on language tasks rather than vision. Specifically, we used word embeddings to predict similarity ratings between pairs of words using the SimLex-999 dataset (Hill, Reichart, & Korhonen, 2015). We used the SimDR modeling framework for this task but used word embeddings as input instead of CNN representations. We used the Word2Vec (Mikolov, Chen, Corrado & Dean, 2013) embeddings as well as the GloVe (Pennington, Socher, & Manning, 2014) embeddings as inputs to the model. We show in Fig. 13 that as the size of the bottleneck layer increases, the prediction performance of our model increases until the full dimensionality of word embeddings (300). This indicates that, unlike CNN representations, word embeddings cannot be compressed into low-dimensional spaces that are still predictive of similarity ratings. This result suggests

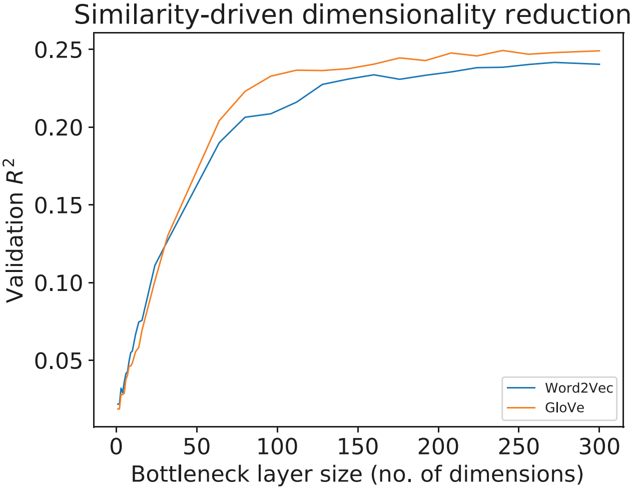


Fig. 13. Explained variance (R^2) of our model in predicting word similarity ratings using GloVe and Word2Vec embeddings at varying bottleneck sizes. The explained variance curve saturates at a much higher value (50–100 dimensions) than that for image representations. Hence, word embeddings cannot be compressed further into low-dimensional representations while capturing human similarity judgments.

that the low dimensionality of extracted psychological representations is a property of the CNN representations as opposed to being a consequence of the supervised dimensionality reduction itself.

6. Discussion

CNNs are popular in the field of computer vision because of their ability to replicate or exceed human-level performance in several visual tasks. CNN representations of natural images are also used as components of cognitive models. We developed a framework for transforming CNN representations to lower dimensions—where interpretation is far less cumbersome—while still being predictive of human behavior. Our findings are based on two cognitive tasks: similarity judgments and categorization. Through this framework, we have shown that only a few dimensions are sufficient to predict human behavior. This finding has implications for two major directions in cognitive science. First, the deep feature sets that are currently being used in both cognitive modeling (Ma & Peters, 2020) and neuroscience (Cichy & Kaiser, 2019; Kriegeskorte, 2015; Kietzmann, McClure, & Kriegeskorte, 2019) are much higher dimensional. Our results suggest that a far small smaller projection of these high-dimensional representations is sufficient for capturing human behavior. Second, our work reinforces the idea of human similarity judgments being able to be captured using relatively few features of images (as also recently shown by Hebart, Zheng, Pereira, & Baker, 2020) and extends this further by demonstrating the same for human categorization.

Furthermore, the low-dimensional spaces identified by our approach can also be used to visualize individual dimensions and show that they code for unique concepts. Hence, they provide insight into potential factors that influence human representations of visual stimuli. We find that as we increase the dimensionality of our model's representations, these representations capture finer levels of distinctions. Since we also find that categorization behavior can also be encoded in a small number of dimensions (on the order of ~ 10 dimensions), our results suggest that CNN representations contain large numbers of features that are redundant for predicting human behavior. Our findings are, however, limited to visual stimuli and do not take into account other forms of perception that may shape people's representations of objects. Future work can leverage richer and multimodal stimuli (such as videos, three-dimensional models, and words) to study the dimensionality of human representations.

We further probed these observations using a set of control studies, each designed to control for a specific factor. Through these studies, we showed that our results are neither a consequence of dataset size nor the modeling approach, but rather of the CNN representations themselves. Our control study on word similarity judgments opens the door for future research on understanding if the number of psychological dimensions required for language are inherently different for language as compared to vision, or if word embeddings from deep networks such as transformers (Devlin, Chang, Lee, & Toutanova, 2019) exhibit the same low dimensionality as that exhibited by CNN representations. Further, the fact that low-dimensional representations are sufficient to predict similarity judgments derived from CNN representations themselves as well as for standard image classification, suggests that CNN representations might inherently be compressible for perceptual tasks. This means that while typical CNNs have thousands of dimensions in their embedding layers only a small subset of these are useful. These findings have far-reaching implications for the field of computer vision; future work should evaluate whether it also holds on larger datasets. Another direction for future investigation could entail analyzing the complementary set of dimensions to get a better handle on why only a small set of dimensions are sufficient to perform standard visual tasks. These other dimensions could contain low-level visual information, such as the orientation or color of an image, which are not essential for performing high-level visual tasks. Finally, while analyzing similarity judgments is a typical and widely used task for studying human representations, it is possible that these other features are a component of psychological representations that are not tapped by similarity judgments, suggesting that exploring a wider range of tasks may be another valuable direction for future research.

Acknowledgments

A preliminary version of the analysis of similarity judgments was presented at the Annual Conference of the Cognitive Science Society and appears in the proceedings of that conference (Jha, Peterson, & Griffiths, 2020). This work was supported by grants 1718550 and 1932035 from the National Science Foundation. AJ was supported by a Google Ph.D. fellowship in computational neural and cognitive sciences.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 409–429.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, *11*(1), 5418.

Bilenko, N. Y., & Gallant, J. L. (2016). Pyrcca: Regularized kernel canonical correlation analysis in Python and its applications to neuroimaging. *Frontiers in Neuroinformatics*, *10*, 49.

Chun-Hsiao Yeh, Y. C. (2022). IN100pytorch: Pytorch implementation: Training resnets on imagenet-100. Retrieved from https://github.com/danielchyeh/ImageNet-100-Pytorch.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305–317.

Cohen, G. (2000). Hierarchical models in cognition: Do they have psychological reality? *European Journal of Cognitive Psychology*, *12*, 1–36.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Piscataway, NJ: IEEE.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, Volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, MN: Association for Computational Linguistics.

Gershman, S. J., & Tenenbaum, J. B. (2015). Phrase similarity in humans and machines. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 776–781). Austin, TX: Cognitive Science Society.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778). Piscataway, NJ: IEEE.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185.

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Jha, A., Peterson, J. C., & Griffiths, T. L. (2020). Extracting low-dimensional psychological representations from convolutional neural networks. In *Proceedings of the 42nd annual conference of the Cognitive Science Society* (pp. 2180–. 2186). Austin, TX: Cognitive Science Society.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. *Oxford Research Encyclopedia of Neuroscience*. *25*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264086.013.46

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, Conference Track Proceedings*.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*(1), 417–446.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis— Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., & Weinberger, K. Q., (Eds.), *Advances in neural information processing systems* (Vol. *25*, pp. 1097–1105)

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLOS Computational Biology*, *12*(4), 1–26.

Lake, B., Zaremba, W., Fergus, R., & Gureckis, T. (2015). Deep neural networks predict category typicality ratings for images. In Dale, R., Jennings, C., Maglio, P., Matlock, T., Noelle, D., Warlaumont, A., & Yoshimi, J., (Eds.), *Proceedings of the 37th annual conference of the Cognitive Science Society* (pp. 1243–1248). Austin, TX: Cognitive Science Society

Luce, R. D. (1959). *Individual choice behavior*. Oxford, England: John Wiley.

Ma, W. J., & Peters, B. (2020). *A neural network walks into a lab: Towards using deep nets as models for human behavior* (Preprint). arXiv:2005.02181.

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

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (Eds.), *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, Workshop Track Proceedings*.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-it object representation. *Frontiers in Psychology*, *4*, 128.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Stroudsburg, PA: Association for Computational Linguistics.

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.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., & Russakovsky, O. (2019). Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 9617–9626).

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rumelhart, D. E., & Todd, P. M. (1993). *Learning and connectionist representations* (pp. 3–30). Cambridge, MA: MIT Press.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Sanders, C. A., & Nosofsky, R. M. (2020). Training deep networks to construct a psychological feature space for a natural-object category domain. *Computational Brain & Behavior*, *3*(3), 229–251.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.

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

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Y. Bengio & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, Conference Track Proceedings*.

Singh, P., Peterson, J. C., Battleday, R. M., & Griffiths, T. L. (2020). End-to-end deep prototype and exemplar models for predicting human behavior. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. (pp. 634–640). Austin, TX: Cognitive Science Society.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Material

APPENDIX A

A.1Image embeddings along top principal components

In Fig. 4, we visualized image embeddings for three datasets by projecting them along the top principal components of the low-dimensional representations learned by SimDR for one validation fold (out of six folds in all). We found that these dimensions encode interpretable concepts. Here, we visualize image embeddings projected along the top principal components for three other validation folds, but for the same datasets (*animals*, *various*, and *vehicles*), to substantiate our claim that these concepts are captured consistently across all cross-validation folds. In the case of the *animals* dataset, we find that the first dimension consistently captures non-mammals versus mammals across all folds shown in Fig. A1. Similarly, the first dimension of the *various* dataset goes from inanimate objects to living beings and their parts (Fig. A2). Finally, PC1 of the *vehicles* dataset goes from manually operated or two-wheeler vehicles to trucks and buses, as shown in Fig. A3.

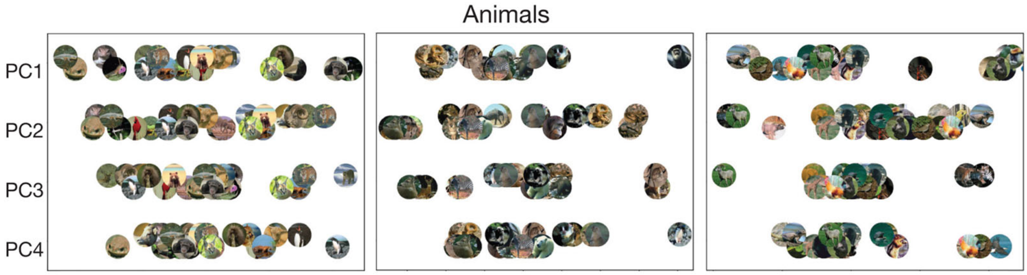


Fig. A1. Image embeddings visualized along the top four principal components of low-dimensional SimDR representations learned on the *animals* dataset. Each plot corresponds to a different validation set (on a 6-fold CV analysis), and consistently shows a distinction between mammals and non-mammals along PC1.

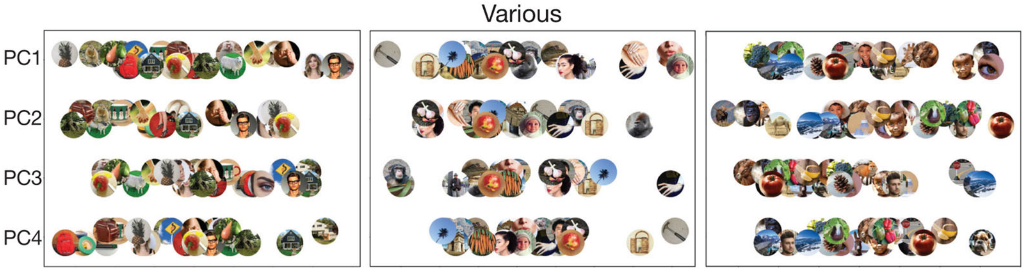


Fig. A2. Image embeddings visualized along the top four principal components of low-dimensional SimDR representations learned on the *various* dataset. Each plot corresponds to a different validation set (on a 6-fold CV analysis). The first PC in all cases separates inanimate objects from living beings.

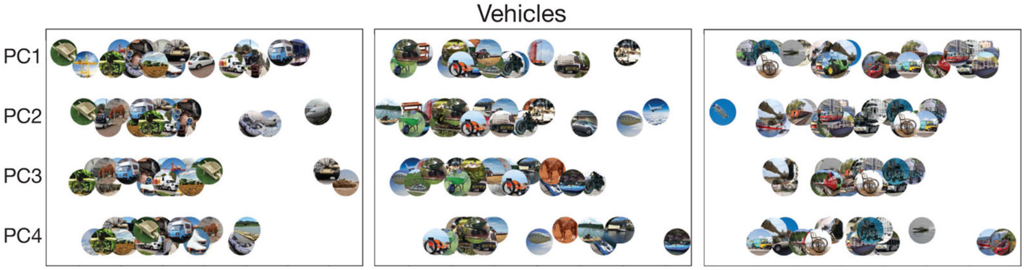


Fig. A3. Image embeddings visualized along the top four principal components of low-dimensional SimDR representations learned on the *vehicles* dataset. Each plot corresponds to a different validation set (on a six-fold CV analysis), where the first PC consistently goes from manually operated or two-wheeler vehicles to buses and trucks.