Extracting low-dimensional psychological representations from convolutional neural networks

Aditi Jha[1] **(aditijha@princeton.edu)**
Joshua Peterson[2] **(joshuacp@princeton.edu)**
Thomas L. Griffiths[2,3] **(tomg@princeton.edu)**
[1]Department of Electrical Engineering
[2]Department of Computer Science
[3]Department of Psychology
Princeton University

Abstract

Deep neural networks are increasingly being used in cognitive modeling as a means of deriving representations for complex stimuli such as images. While the predictive power of these networks is high, it is often not clear whether they also offer useful explanations of the task at hand. Convolutional neural network representations have been shown to be predictive of human similarity judgments for images after appropriate adaptation. However, these high-dimensional representations are difficult to interpret. Here we present a method for reducing these representations to a low-dimensional space which is still predictive of similarity judgments. We show that these low-dimensional representations also provide insightful explanations of factors underlying human similarity judgments.

Keywords: similarity judgments; neural networks; deep learning; dimensionality reduction; interpretability

Introduction

Judging similarity between any pair of stimuli is an ambiguous problem: deciding what counts as similar is subjective and sensitive to context (Medin, Goldstone, & Gentner, 1993). Nevertheless, people are relatively consistent in making similarity judgments, which is perhaps explained in part by the biases they develop towards emphasizing some stimulus features over others (*e.g.,* shape over size, color, or material; Diesendruck & Bloom, 2003). Understanding the features (and the weights upon them) that people employ when evaluating the similarity of complex stimuli like images remains an open problem.

Deep neural networks have been demonstrated to be predictive of multiple aspects of human visual perception in visuoperceptual tasks (*e.g.,* Lake, Zaremba, Fergus, & Gureckis, 2015; Kubilius, Bracci, & Op de Beeck, 2016). This utility has led to their increasing use as proxies of human cognition to understand mechanisms underlying cognitive processes or as proofs-of-concept to establish the possibility of a certain cognitive strategy (Kriegeskorte, 2015; Cichy & Kaiser, 2019). For example, Sanders and Nosofsky (2020) show that CNNs can be trained using multidimensional scaling representations to derive psychological representations of images. In other work, Peterson, Abbott, and Griffiths (2018) show correspondences between similarities in convolutional neural net (CNN) representations and human similarity judgments for natural images. They find 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 success of CNNs in predicting human similarity judgments suggests that they might also be helpful in identifying the features which inform those judgments. However, CNN representations are high-dimensional, potentially redundant, and likely include psychologically irrelevant stimulus information. An important question, given their increasing use in cognitive modeling is how many relevant features/dimensions they really contribute and what the nature of those features might be.

In this work, we propose a method inspired by previous work by Rumelhart and Todd (1993) which we call similarity-driven dimensionality reduction (SimDR), which obtains low-dimensional projections of CNN image representations that best capture human similarity judgments. Surprisingly, our method reveals that human similarity judgments continue to be well-preserved even up to two orders of magnitude fewer dimensions than previous work. This suggests that the dimensionality of psychological representations is considerably less than the full set of CNN features. We further probe the individual dimensions of these representations that capture concepts essential to judging similarity, and find that most of them are interpretable. In particular, we show that broad categories are given more importance by our model than finer ones captured in subsequent dimensions, in line with the hierarchical structure oft-found to characterize human cognition (Cohen, 2000; Rogers & McClelland, 2004).

Method

Peterson et al. (2018) show that the final representation layer of a CNN can be adapted to better predict human similarity judgments. The size of the final representation in CNNs is typically of the order of 10^3, which makes interpretation difficult. To serve our purpose of leveraging CNN representations to understand human similarity judgments, we require representations that are interpretable and can give us insights into the actual cognitive task.

Rumelhart and Todd (1993) constructed a connectionist model to mimic human similarity judgments. The model takes two stimuli as input and outputs a similarity judgment. The hidden layer is of lower dimensionality than the input, resulting in a compressed representation. Extending this idea to modern CNNs, our method (SimDR) reduces the CNN representations of images to a low-dimensional space which is optimal for predicting human similarity judgments. If the

obtained representations have sufficiently low dimensionality, we can interpret individual dimensions to see what they capture and make inferences about similarity judgments in humans. This model and the data used are explained in the following sections.

Datasets

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"). Each of these sets comprises pairwise similarity ratings from ten people for 120 images, which we employ in all analyses that follow.

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 VGG-19 representations for all our experiments. The task of obtaining low-dimensional representations of images which capture factors underlying human similarity judgments is split by SimDR into two objectives: (a) projecting VGG-19 representations to a low-dimensional space, and (b) predicting human similarity judgments using the low-dimensional representations.

These two objectives are jointly optimized leading to low-dimensional representations that are predictive of human similarity judgments. VGG-19 representations of two input images are passed through a single linear layer of small width (*i.e.,* a bottleneck layer) which projects them to a lower-dimensional space. This is followed by an inner product of the outputs of the bottleneck layer to obtain the predicted similarity rating for the input pair (Fig. 1). 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). For both input images, the weights of the bottleneck layer are shared. The weights 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. This contrasts with the method of Peterson et al. (2018), which learns weights for each of the 4096 input dimensions, or principal component analysis (PCA), which preserves as much information as possible as opposed to just that which is relevant to human similarity judgments (and thus may inflate the estimated intrinsic dimensionality).

We first trained a separate model for each dataset. CNN feature vectors were first normalized such that their norms were one. We used mean squared error loss with L2 regularization to train each model. The L2 coefficient was selected between 10^{-3} and 10^{3} by 6-fold cross-validation over the 120 images. Further, for every dataset, the number of nodes in the bottleneck layer is varied in the range of $1-64$. We also compare the above with a simple unsupervised baseline that alternatively obtains low-dimensional representations by running PCA over the input VGG-19 representations. These low-dimensional representations are then transformed by ridge regression using the method of Peterson et al. (2018) to predict similarity ratings. As above, we vary the number of principal components in the range of $1-64$.

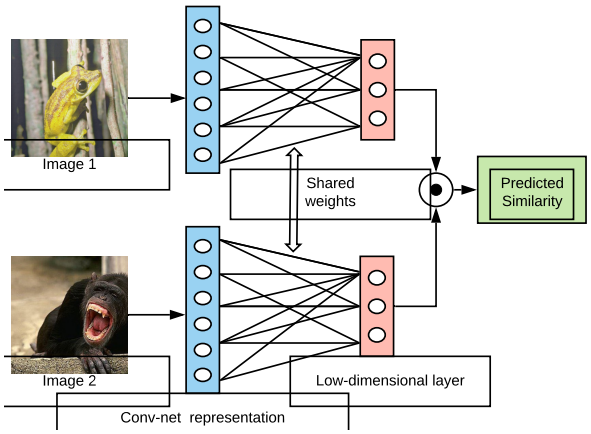

Figure 1: Overview of SimDR. CNN representations for an image pair are down-projected using a shared low-dimensional bottleneck layer. An inner product of the outputs gives predicted similarity rating for the input pair.

Few dimensions predict similarity judgments

We observe for all datasets that the SimDR R^2 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 6 nodes to achieve an R^2 score of 0.6 while the *various* dataset achieves an R^2 of 0.49 at 6 nodes. 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-

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

Table 1: R^2 scores for all datasets (SimDR values are for bottleneck layer of size 64).

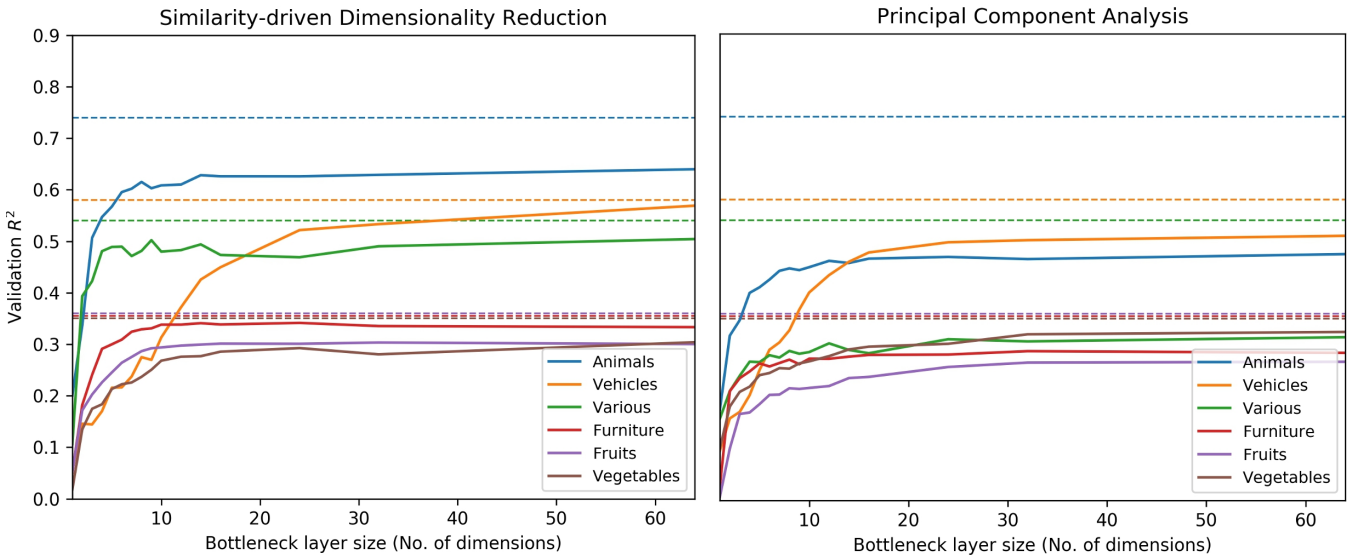

Figure 2: Explained variance (R^2) of our models in predicting human similarity judgments on each dataset. The dashed lines correspond to the prediction performance in Peterson et al. (2018) when all input dimensions are used.

dimensional than CNN representations. Additional evidence for this can be seen in the intrinsic dimensionality of the CNN representations themselves without respect to human judgements. 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 *various*), apparently much unlike humans (Fig. 2).

Interpretation of low-dimensional features

Now that we have demonstrated the sufficiency of low-dimensional representations to predict similarity judgments, we can attempt to interpret the reduced dimensions. For this experiment, we focus on the top 3 datasets based on R^2 score—*animals*, *vehicles*, *various*. As mentioned above, SimDR achieves an R^2 score of 0.6 on the *animals* dataset using only 6 dimensions. On the *various* dataset, it achieves an R^2 score of 0.49 using 6 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 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.

Visualizing individual dimensions

The ability of the low-dimensional representations to predict similarity indicates that they are efficiently encoding information essential for making similarity judgments. Hence, they can further be leveraged to understand what factors allow them to predict similarity judgments. To this end, for each of the three datasets, we visualize image embeddings along the top four principal dimensions of the low-dimensional features learned via SimDR. We visualize validation images for a single fold (out of the 6 cross validation folds), though we observe that the dimensions were consistent across all folds in terms of capturing the same concepts (Fig. 4).

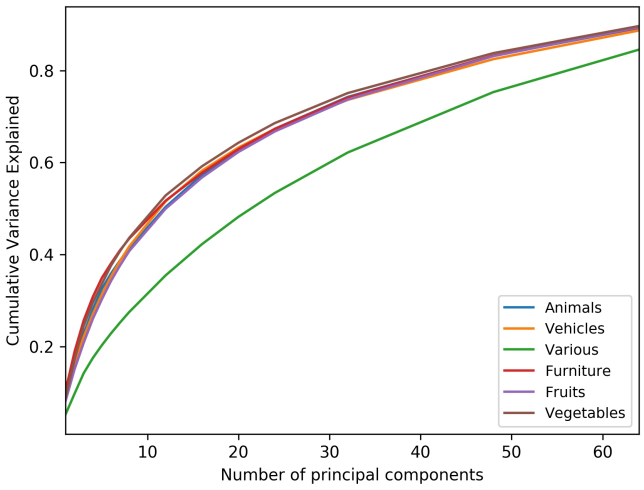

Figure 3: Cumulative variance explained in the full VGG-19 representations as a function of principal component.

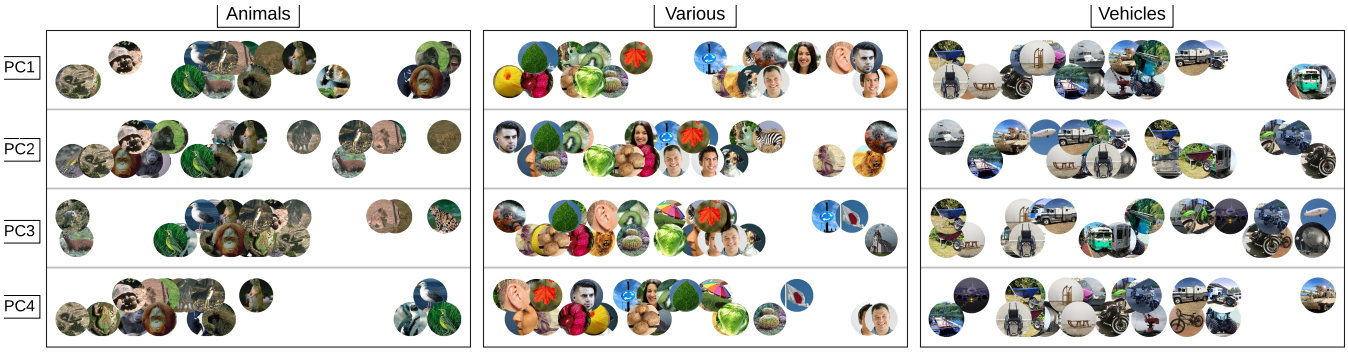

Figure 4: Image embeddings along the top four principal components of low-dimensional SimDR representations.

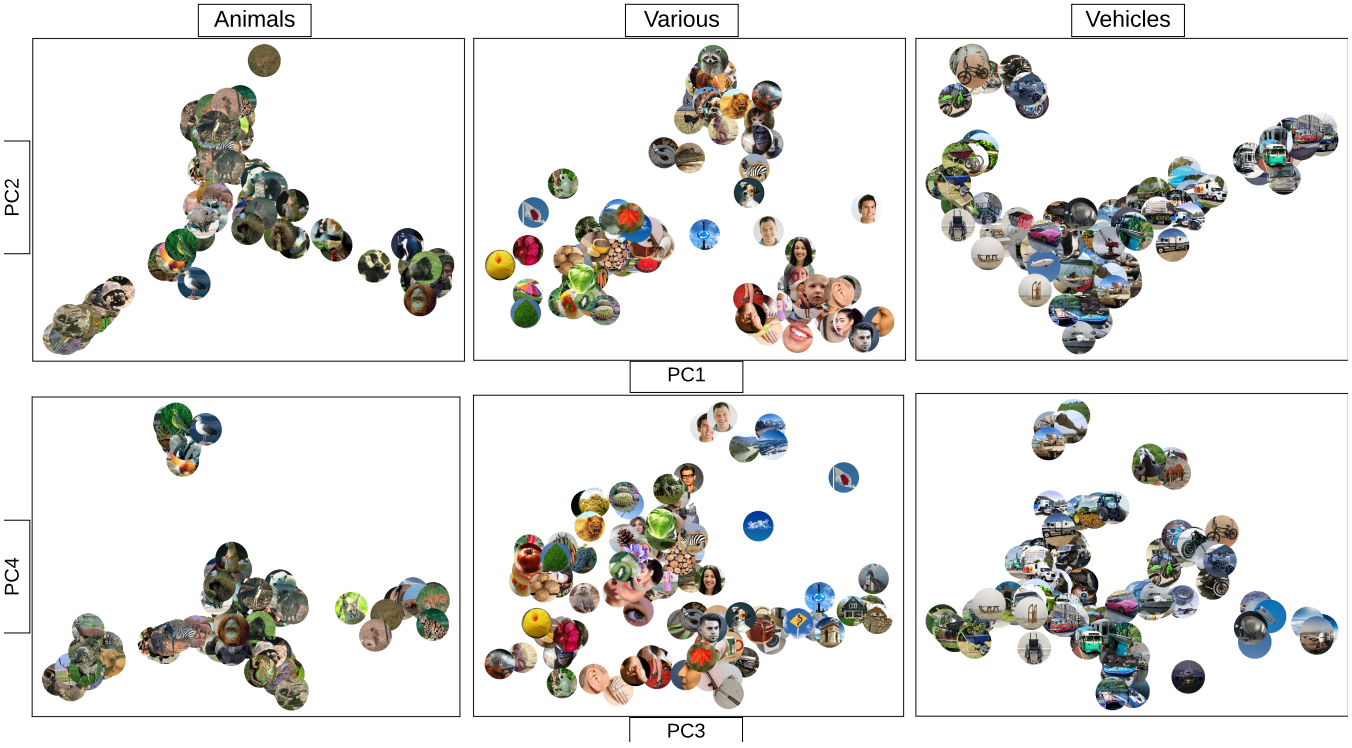

Figure 5: Examples of image embeddings for three datasets using the top principal components of the SimDR representations.

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. 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.

Clusters formed by pairs of dimensions

As an alternative visualization strategy, we explore 2D 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 kind of images are considered similar by the model.

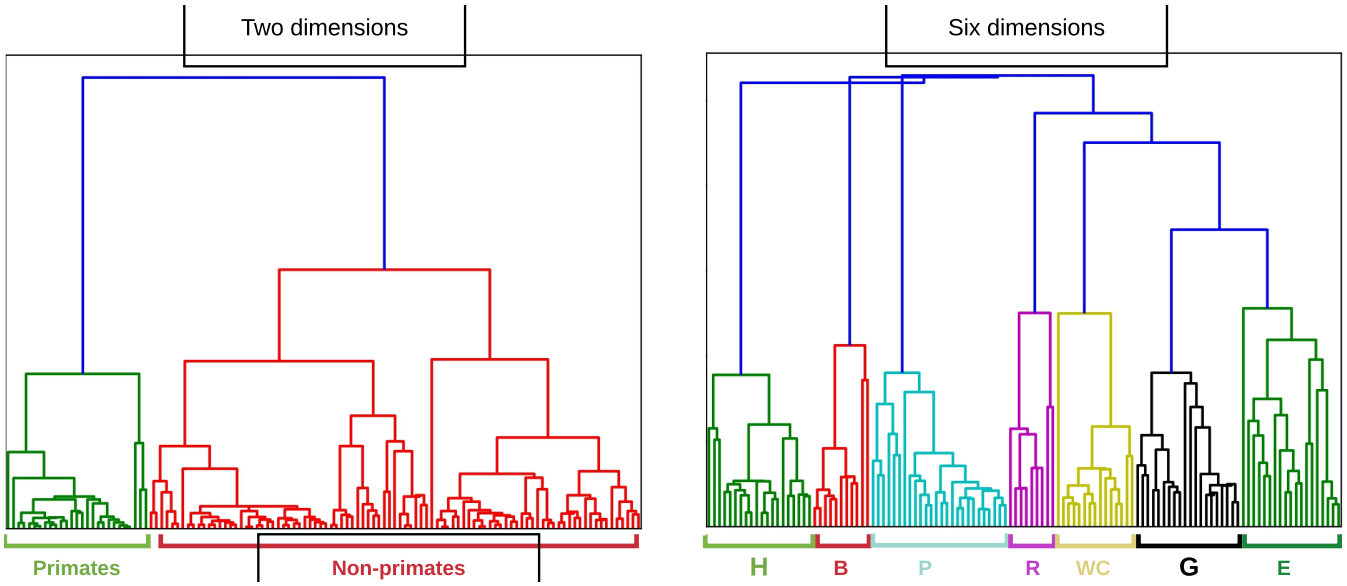

Figure 6: Dendrograms of hierarchical clustering for 2-dimensional representations and 6-dimensional representations on *animals* dataset. *H: Herps, B: Birds, P: Primates, R: Rodents, WC: Wild cats, G: Grazers, E: Dogs, Bears and Large animals.*

Echoing Peterson et al. (2018), we observe clusters for herptiles, 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.

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 dendrograms (Shepard, 1980) for the *animals* dataset. Fig. 6 shows that when the size of the bottleneck layer in SimDR is 2, 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, however, 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 6-dimensional representations is closely related to that formed using human similarity data in Peterson et al. (2018).

We observe that increasing the bottleneck width introduces further categorical distinctions in other datasets too. For the *various* dataset, at a bottleneck width of 4, we observe distinct clusters for animals and humans (and their body parts). In the case of the *vehicles* dataset, 4-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 organisation in neural networks (Saxe, McClelland, & Ganguli, 2019).

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?

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 6 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 model trained on fruits and the model trained on vegetables have overlapping latent factors and hence, their similarity predictions are also based on some common factors. The same is true for *animals* and *vehicles* datasets. While it seems reasonable for *fruits* and *vegetables* to share

Figure 7: Inter-domain relatedness (R^2) as measured by regularized CCA between all domain pairs.

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).

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 6-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, 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. 9 shows the performance of the pooled model in 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*).

Conclusion

Our work shows that CNN representations can be transformed to lower dimensions—where interpretation is far less cumbersome—while still being predictive of similarity judgments. We also observe that only a few dimensions are required to predict psychological representations as opposed to a considerably larger, full set of CNN features. This finding is interesting because the deep feature sets increasingly being used in both cognitive modeling (for a review, see Ma & Peters, 2020) and neuroscience (Kriegeskorte, 2015; Kietzmann, McClure, & Kriegeskorte, 2019; Cichy & Kaiser, 2019) are much higher-dimensional. Indeed, some work may already suggest that our findings could generalize to modeling neural activity as well (Mur et al., 2013), though future work must bear this out.

Moreover, in this low-dimensional space, we are able to visualize individual dimensions and show that they code for unique concepts. Hence, they provide insights into potential factors that influence human similarity judgments, and potentially various other visual tasks. We observe that in-

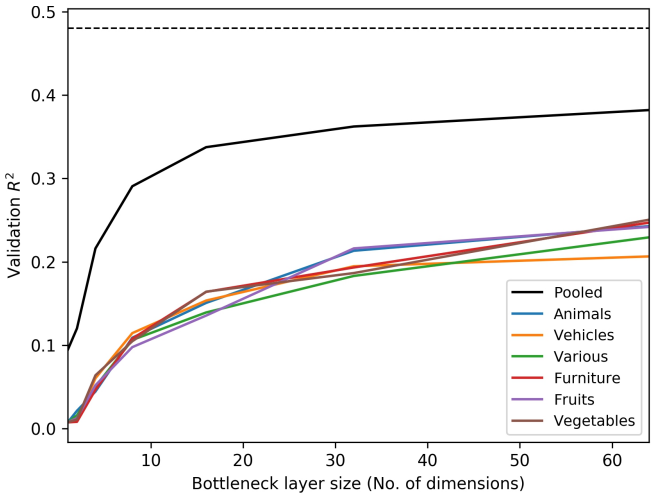

Figure 8: Performance of models tested on all domains (with varying bottleneck layer size). The dashed line shows the performance of the model trained on all domains in Peterson et al., 2018. Solid lines correspond to models trained on different datasets.

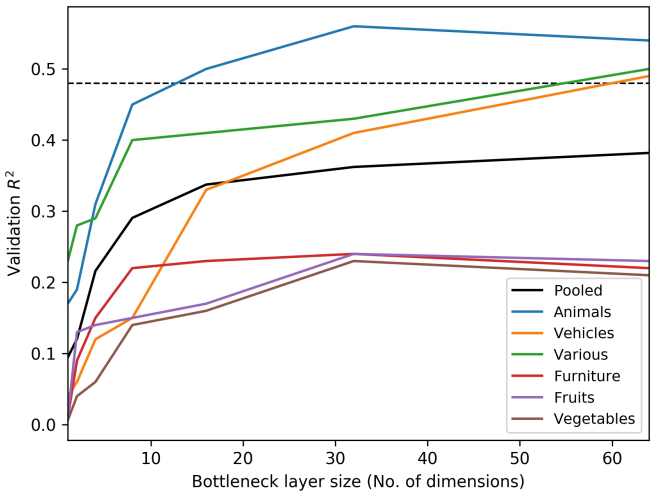

Figure 9: Performance of pooled model tested on individual domains and on all domains (with varying bottleneck layer size). The dashed line shows the performance of the model trained on all domains in Peterson et al., 2018. Solid lines correspond to the pooled model tested on different datasets.

creasing the size of the bottleneck layer introduces finer levels of distinction, mirroring hierarchical clustering in human cognition. These results together show the ability of CNN representations to both predict and explain human similarity judgments using a few dimensions.

This work takes a step towards showing that psychological representations can be predicted by far fewer dimensions than used in CNNs; and that they are not only quantitatively predictive of human similarity judgments but provide insights about how people make similarity judgments. We think our approach can help bridge the interpretation gap between CNN representations and psychological representations by providing interpretable factors which influence human similarity judgment.

Acknowledgements

This work was supported by the National Science Foundation (grant number 1718550), and the School of Engineering and Applied Sciences at Princeton University.

References

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. doi: 10.3389/fninf.2016.00049

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305 - 317. doi: https://doi.org/10.1016/j.tics.2019.01.009

Cohen, G. (2000, 03). Hierarchical models in cognition: Do they have psychological reality? *European Journal of Cognitive Psychology*, *12*, 1-36. doi: 10.1080/095414400382181

Diesendruck, G., & Bloom, P. (2003). How specific is the shape bias? *Child Development*, *74*(1), 168-178.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019, 01). Deep neural networks in computational neuroscience. *Oxford Research Encyclopedia of Neuroscience*.

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.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016, 04). 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 R. Dale et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

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

Medin, D., Goldstone, R., & Gentner, D. (1993, 04). Respects for similarity. *Psychological Review*, *100*, 254-278. doi: 10.1037/0033-295X.100.2.254

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.

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.

Rogers, T. T., & McClelland, J. L. (2004). Semantic cognition: A parallel distributed processing approach..

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In *Attention and performance xiv (silver jubilee volume): Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (p. 330). Cambridge, MA, USA: MIT Press.

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*. doi: https://doi.org/10.1007/s42113-020-00073-z

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. doi: 10.1073/pnas.1820226116

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 *International Conference on Learning Representations.*