Adapting Deep Network Features to Capture Psychological Representations: An Abridged Report

Joshua C. Peterson
UC Berkeley
peterson.c.joshua@gmail.com

Joshua T. Abbott
UC Berkeley
joshua.abbott@berkeley.edu

Thomas L. Griffiths
UC Berkeley
thomas_griffiths@berkeley.edu

Abstract

Deep neural networks have become increasingly successful at solving classic perception problems (e.g., recognizing objects), often reaching or surpassing human-level accuracy. In this abridged report of Peterson *et al.* [2016], we examine the relationship between the image representations learned by these networks and those of humans. We find that deep features learned in service of object classification account for a significant amount of the variance in human similarity judgments for a set of animal images. However, these features do not appear to capture some key qualitative aspects of human representations. To close this gap, we present a method for adapting deep features to align with human similarity judgments, resulting in image representations that can potentially be used to extend the scope of psychological experiments and inform human-centric AI.

1 Introduction

The resurgence of neural networks in the form of *deep learning* has continued to dominate object recognition benchmarks in the field of computer vision, often attaining near or above human-level accuracy for a variety of perceptual tasks, most notably through recent advances in classifying thousands of objects within natural images [Krizhevsky *et al.*, 2012; He *et al.*, 2015]. Part of the success of these models is due to their ability to learn effective feature representations of high-dimensional inputs (e.g., complex color images); a challenge that human perception must also confront [Austerweil and Griffiths, 2013]. As a result, cognitive scientists have started to explore how the representations learned by these networks can be used in models of human behavior for perceptual tasks such as predicting the memorability of objects in images [Dubey *et al.*, 2015] and predicting judgments of category typicality [Lake *et al.*, 2015].

While deep learning models continue to mimic a growing list of human-like abilities, a number of core questions remain unanswered about the relevance of these models to actual human cognition and perception. For instance, features of the input learned using these networks excel in predicting certain human judgments, but how are these feature representations related to human psychological representations? For decades, psychologists have studied the underlying structures that support mental representations such as geometric spaces and hierarchies [Shepard, 1980] that are known to aid crucial learning and inference strategies [Griffiths *et al.*, 2010; Tenenbaum *et al.*, 2011]. For this reason, one should expect any satisfactorily human-like representation to mirror these structures. At first glance, it would seem that the ability of these representations to predict typicality judgments and stimulus memorability would constitute robust evidence of their relevance to people, however recent work has shown that neural networks that classify images can be systematically deceived by imperceptible image transformations [Szegedy *et al.*, 2013], casting doubt on their similarity to humans.

Understanding the relationship between the representations found by deep learning and those of humans is an important question in cognitive science. Simply having a good approximation to how people represent images would allow cognitive scientists to test psychological theories using complex, realistic stimuli. Indeed, tasks such as creating stimulus sets that uniformly span psychological space are far from trivial. In addition, since human generalization and categorization behavior is still the standard for solving such problems in artificially intelligent systems, it is imperative to understand where the two diverge.

In this abridged report of Peterson *et al.* [2016], we address this question directly by examining how well features extracted from state-of-the-art deep neural networks predict human similarity judgments. An initial evaluation shows that these features account for a significant amount of variance in human judgments, but fail to capture qualitative distinctions that are key to human representations. We then develop a method for adapting deep network features to better predict human similarity judgments, and show that this approach can reproduce those qualitative distinctions. These results suggest that while raw features produced by deep learning may not be suitable for use in modeling cognition, they can be modified to bring them into close alignment with human representations.

2 Deep Representations

In general, deep neural networks (DNNs) are neural networks that have depth in terms of their number of hidden layers between input and output [Bengio, 2009]. In the past few years, training such networks to understand aspects of large, complex data sets has led to a number of advances in vision and language applications [LeCun *et al.*, 2015].

In computer vision, the majority of this progress has been driven by a particular DNN called a convolutional neural network (CNN) [LeCun *et al.*, 1989]. CNNs get their name from the use of convolutional layers, which learn a set of image filters that produce feature maps of spatially-organized inputs like images. This allows for a drastic decrease in the number of parameters the network must learn, which would otherwise explode exponentially in a fully connected network with high-dimensional inputs. The typical CNN architecture includes a series of hidden convolutional layers, followed by a smaller number of fully connected layers, and finally a layer that generates the final output or classification. While CNNs were initially developed over two decades ago, they came to mainstream popularity in 2012 when a 7-layer architecture named AlexNet [Krizhevsky *et al.*, 2012] won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), reducing the previous winner's error rate by an uncommonly large margin. Since then, a deeper CNN has won the contest every year, currently dominated by Microsoft's 150-layer network which obtained a best-of-top-5 error rate of 4.94%, surpassing the accuracy of non-expert humans at 5.1% [He *et al.*, 2015].

Interestingly, CNNs produce much more than just their outputs (e.g., a category label for an image); they can also return feature representations at each layer of the network. The "deep representations" learned by these networks have proven useful in predicting human behavior. Dubey *et al.* [2015] used representations extracted from the last fully-connected layer of a CNN to predict the intrinsic memorability of objects. That is, the objects that humans are jointly likely to remember or forget in a large complex natural scene database. The correlation between estimates of memorability and the original memorability scores for each object matched human consistency (i.e. the correlation between memorability scores of random splits of the full sample of subjects). Similarly, Lake *et al.* [2015] were able to reliably predict human typicality ratings of eight object categories using the same network and features, and called for cognitive scientists to pay attention to deep learning since categorization is a foundational problem in the field.

Deep representations are also beginning to interest the neuroscience community. For example, CNN activations have been used to predict monkey inferotemporal (IT) cortex activity [Yamins *et al.*, 2014], as well as both low- and high-level activity in human visual areas [Agrawal *et al.*, 2014]. Delving deeper, Khaligh-Razavi and Kriegeskorte [2014] found that a CNN best explained IT cortex representations out of a set of 37 well-known models from both the computer vision and neuroscience fields, although no model completely explained all of the variance, unsupervised models being the worst of all of them.

Although CNN representations currently do the best job of predicting neural activity as measured by Blood Oxygenation Level Dependent (BOLD) response, this does not guarantee that we can explain psychological representations as a result. In fact, Mur *et al.* [2013] was partly successful in predicting human similarity judgments (a classic index of psychological representations) from IT cortex representations. However, the key categorical distinctions in the human representations were not well predicted: human IT cortex representations were more similar to monkey IT cortex representations than they were to human psychological representations. In the remainder of the paper, we use a similar approach to evaluate how well deep network features align with human psychological representations, and to explore how the correspondence between the two can be increased.

3 Evaluating Representations

Our first step is to evaluate the potential correspondence between deep network features and psychological representations. Unlike neural representations, psychological representations cannot be measured directly. However, both spatial and hierarchical psychological representations for N objects can be recovered given an $N \times N$ matrix of similarity judgments using methods such as multidimensional scaling and hierarchical clustering [Shepard, 1980]. We thus reduce the problem to one of capturing human similarity judgments, subjecting both human judgments and model predictions to these different methods of extracting representations. We approach this problem by taking the inner-product of the deep feature representations of each pair of images (a measure of similarity between two vectors). We then compute the correlation between these pairwise vector similarities and human similarity judgments for the same stimulus pairs, which gives us a measure of the correspondence we want to evaluate.

Behavioral Experiment. We collected pairwise similarity ratings for 120 color animal photographs (examples shown in Figure 1) through Amazon Mechanical Turk. Participants were initially shown 8 diverse examples to help prevent bias due to the sampling of the pairs, and were then instructed to rate the similarity of four pairs of animal images on a scale from 0 (not similar at all) to 10 (very similar). Workers could repeat the task with new pairs as many times as they wanted. There were $7,140$ possible comparisons, each of which we ensured was rated by 10 unique participants, for a total of $71,400$ ratings from 209 different participants. The result was a 120×120 similarity matrix after averaging over judgments.

Feature Extraction. We extracted features for each image in our data set using three different popular *off-the-shelf* CNNs of varying complexity that were pretrained in Caffe [Jia *et al.*, 2014]. Specifically, we used CaffeNet (based on original AlexNet), VGG16 [Simonyan and Zisserman, 2014], and GoogLeNet [Szegedy *et al.*, 2014], the layer depths of which were 7, 16, and 22 respectively. GoogLeNet and VGG16 achieve roughly half the error rates of AlexNet. Each network had already been trained to classify 1000 object categories from previous ILSVRC competitions. A feedforward pass of each flattened image vector into each network yields feature responses at each layer. For our analysis, we extracted the last layer of each network before

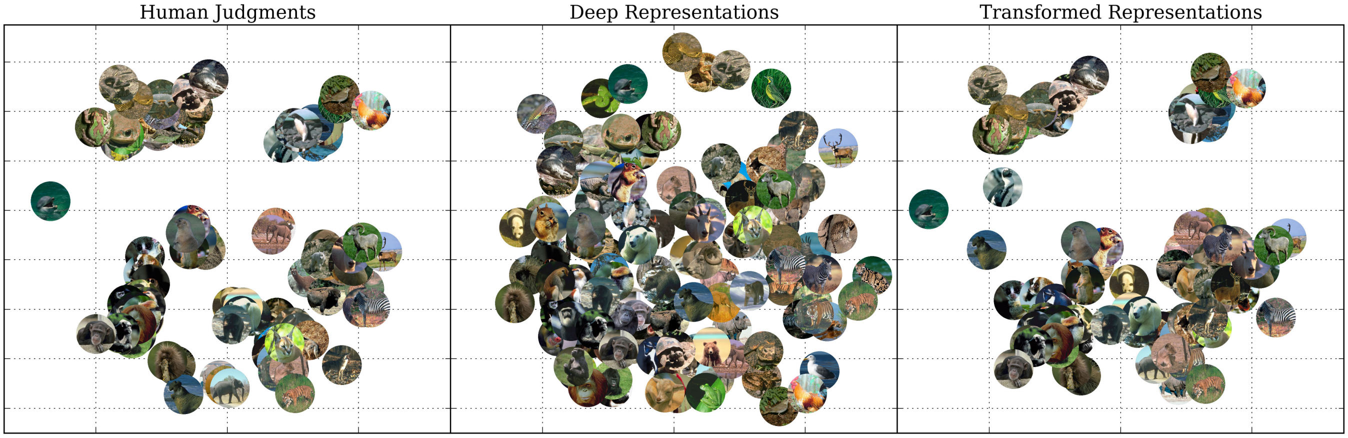

Figure 1: Multidimensional scaling solutions for similarity matrices obtained from human judgments (left), non-transformed deep representations (center), and transformed deep representations (right).

the classification layer. For CaffeNet and VGG16, this is a 4096-dimensional fully-connected layer, while the last layer in GoogleNet is a 1000-dimensional average pooling layer. Lastly, we also extracted Histograms of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) representations for comparison since such features represent the generic representations of choice for tasks in computer vision prior to the popularity of deep learning.

Results. Correlations (R^2) between human and DNN representations for each network were: .32 (CaffeNet), .35 (Google), .43 (VGG), and .008 (HOG+SIFT). Raw representations from all three networks show medium to high correlations with the human data. In general, deeper networks with better ImageNet classification accuracy like GoogLeNet and VGG16 did better than CaffeNet, which is considerbly more shallow. The HOG+SIFT baseline did surprisingly poorly, explaining very little variance as compared to the deep representations, suggesting that while these features are useful for many computer vision tasks, they differ in large part from the representations humans employ when judging animal similarity.

Although the VGG representation explained a fair amount of variance, further analyses revealed that the most crucial structural aspects of the human representations were not preserved. The first and second panels of Figure 1 show multidimensional scaling (MDS) solutions for the original human data and the predictions from the unaltered deep representations. While the structure of the MDS solutions for the predicted judgments looks reasonable (e.g., zebras are next to other zebras), major categorical divisions are not preserved. Hierarchical clusterings of the actual and predicted human judgments (the first and second panels of Figure 2) show a similar pattern of results: human judgments exhibit several major categorical divisions, whereas much of this structure is lost in the predicted data.

4 Adapting Representations

After quantifying the discrepancy between deep and human representations, we can attempt to bring them into closer alignment. First, consider that the final hidden layer feature representation in a neural network can be thought of as the input to a final linear classification layer, such that the problem solved by the final weight matrix is a linear transformation (which is then often scaled by a softmax function to covert to class probabilities). This can be thought of as a rescaling of the final stimulus representation to solve the categorization problem. This suggests that we should not think about the features extracted by the network as a static representation, but as the ingredients for a transformation that solves a problem. Thinking in these terms, we show that we can easily solve for a linear transformation that better captures human similarity judgments.

Similarity Model. Any similarity matrix $\mathbf{S}$ can be decomposed into the matrix product of a feature-by-object matrix $\mathbf{F}$, its transpose, and a diagonal weight matrix $\mathbf{W}$,

$$\mathbf{S} = \mathbf{F}\mathbf{W}\mathbf{F}^T \qquad (1)$$

This formulation is similar to that employed by additive clustering models [Shepard and Arabie, 1979], wherein $\mathbf{F}$ represents a binary feature identity matrix (and is similar to Tversky's (1977) famous model of similarity). Given an existing feature-by-object matrix $\mathbf{F}$, the diagonal of $\mathbf{W}$ can be solved for using linear regression where the predictors for each similarity s_{ij} are the product of the values of each feature for the objects i and j. When $\mathbf{W}$ is the identity matrix, this reduces to the model evaluated in the previous section.

$$s_{ij} = \sum_{i=1}^{N_f} w_k f_{ik} f_{jk}. \qquad (2)$$

This results in a convex optimization problem that can be solved straightforwardly, allowing us to find a transformation of the deep features with a closer correspondence to human similarity judgments.

Analysis. With such a large number of predictors, regularization is critical to avoid overfitting. We used ridge regression ($L2$ regularization) and performed grid search on 6-fold cross-validated generalization performance to find the best

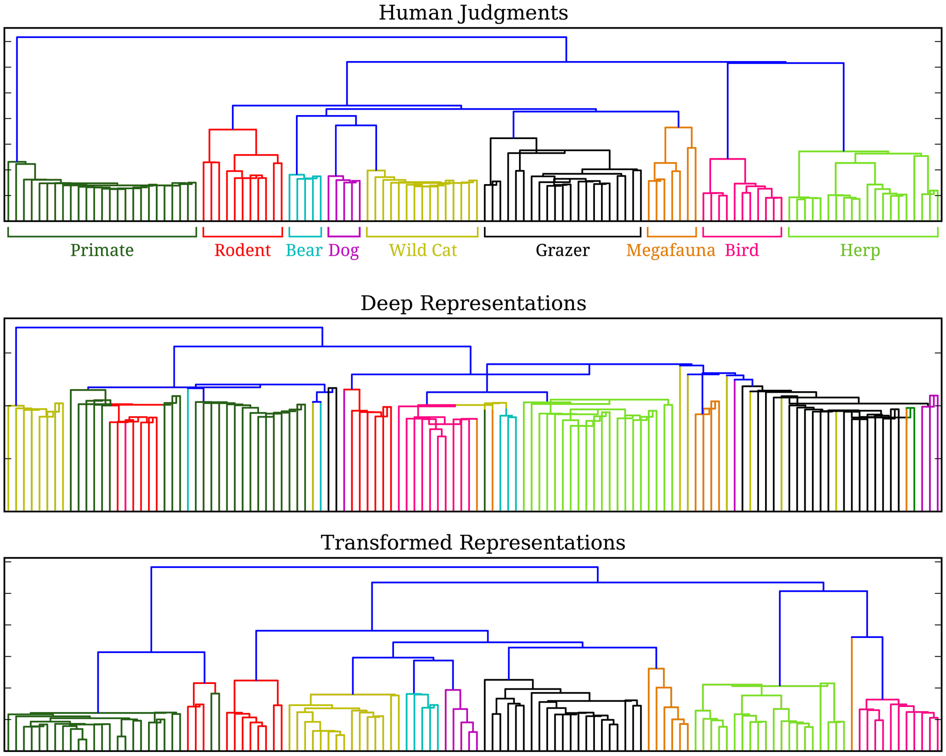

Figure 2: Hierarchical clustering of human judgments (top), deep representations (middle), and transformed representations (bottom). Human judgments resulted in nine interpretable clusters, grouped by color and semantic category label in the top panel. The leaves of the deep and transformed representation clusterings are color-coded relative to the human judgments.

regularization parameter. As an additional control against overfitting, we compared model performance with several baselines. In Baseline 1, we shuffled the rows of the feature matrix. In Baseline 2, the columns of the feature matrix were randomly permuted for each row separately. Lastly, Baseline 3 simply combined the shuffling schemes from the first two baselines. In all three cases, the randomized feature matrices were subjected to the same set of analyses as the true features, allowing us to check for spurious correlations.

Results. Performance scores (average cross-validated R^2) for predicting human similarity judgments using the representations of each network were: .69 (CaffeNet), .72 (Google), .84 (VGG), and .09 (HOG+SIFT). All five models performed considerably well, each showing improvement over the non-weighted models. VGG16 performed best, accounting for 84% of the variance. Training using the estimated regularization parameter on the entire dataset yielded an R^2 of 91%. In contrast, all three baseline models explained essentially no variance ($R^2 < 0.01$), suggesting that our results were not spurious correlations due to the number of features. Crucially, the MDS solution for the improved predictions is almost identical to the original human spatial representation. The same improvements were found in hierarchical clusterings of actual and predicted similarity matrices (1st and 3rd panels of Figure 2), this time largely in the form of top-level parent nodes.

5 Discussion

The current work constitutes the first formal comparison of deep representations to human psychological representations. Initial results using currently high-performing CNN classifiers show that the two representations are moderately correlated, but diverge in terms of crucial structural characteristics, a problem exhibited by similar experiments using neural representations as opposed to deep features [Mur *et al.*, 2013]. Our method of overcoming this problem appears to have been largely successful: human representations were almost completely reconstructed by our adjusted CNN features. Using features extracted from deep CNNs provides an opportunity to estimate psychological representations from raw sensory inputs (e.g. pixels). However, one potential limitation of this work is the generalizability of the transformation learned to broader sets of concepts beyond animal images. Addressing this will require replication and transfer across diverse image datasets at varying taxonomic depths. To the extent that this can be established, we envision our method as a standard tool for studying cognitive processes with natural images by leveraging modern machine learning breakthroughs, and a benchmark for improving non-human systems that take human intelligence as inspiration.

Acknowledgments

This work was supported by grant number FA9550-13-1-0170 from the Air Force Office of Scientific Research. We thank Alex Huth for help with image selection.

References

Pulkit Agrawal, Dustin Stansbury, Jitendra Malik, and Jack L. Gallant. Pixels to Voxels: Modeling Visual Representation in the Human Brain. *arXiv preprint arXiv:1407.5104*, 2014.

Joseph L. Austerweil and Thomas L. Griffiths. A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review*, 120(4):817, 2013.

Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

Rachit Dubey, Joshua Peterson, Aditya Khosla, Ming-Hsuan Yang, and Bernard Ghanem. What makes an object memorable? In *International Conference on Computer Vision (ICCV)*, 2015.

Thomas L Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B Tenenbaum. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput Biol*, 10(11), 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Brenden M Lake, Wojciech Zaremba, Rob Fergus, and Todd M. Gureckis. Deep neural networks predict category typicality ratings for images. In *Proceedings of the 37th Annual Cognitive Science Society*, 2015.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Marieke Mur, Mirjam Meys, Jerzy Bodurka, Rainer Goebel, Peter A. Bandettini, and Nikolaus Kriegeskorte. Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, 4, 2013.

Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Adapting deep network features to capture psychological representations. *arXiv preprint arXiv:1608.02164*, 2016.

Roger N Shepard and Phipps Arabie. Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86(2):87, 1979.

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

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.