Can Deep Neural Networks Rival Human Ability to Generalize in Core Object Recognition?

Jonas Kubilius[*,1,3] (qbilius@mit.edu), Kohitij Kar[*,1,2] (kohitij@mit.edu),
Kailyn Schmidt[1] (kailyn@mit.edu), James J. DiCarlo[1,2] (dicarlo@mit.edu)
[1]McGovern Institute for Brain Research and [2]Department of Brain and Cognitive Sciences, MIT, Cambridge, Massachusetts
[3]Brain and Cognition, KU Leuven, Leuven, Belgium
[*] These authors contributed equally

Abstract

Humans are thought to transfer their knowledge well to unseen domains. This putative ability to generalize is often juxtaposed against deep neural networks that are believed to be mostly domain-specific. Here we assessed the extent of generalization abilities in humans and ImageNet-trained models along two axes of image variations: perturbations to images (e.g., shuffling, blurring) and changes in representation style (e.g., paintings, cartoons). We found that models often matched or exceeded human performance across most image perturbations, even without being exposed to such perturbations during training. Nonetheless, humans performed better than models when image styles were varied. We thus asked if there was any linear decoder that, when applied on model features, would rectify model performance. By adding examples from all representation styles to decoder training, we found that models matched or surpassed human performance in all tested categories. Our results indicate that ImageNet-trained model encoding space is sufficiently rich to support suprahuman-level performance across multiple visual domains.

Keywords: deep nets; object recognition; generalization

Introduction

Humans are thought to be able to learn rapidly from few examples and flexibly transfer their knowledge to new environments and tasks (Hassabis, Kumaran, Summerfield, & Botvinick, 2017). In everyday visual perception tasks, humans appear to be robust to various visual perturbations such as blur and occlusion as well as drastic changes in input statistics, seamlessly switching from recognizing objects in photographs to recognition in line drawings even in the absence of prior experience (Hochberg & Brooks, 1962). In contrast, while hugely successful in reaching high performance in various object recognition challenges, deep neural networks nonetheless show domain-specificity such that changes to input statistics decrease their performance (Kornblith, Shlens, & Le, 2018). But while humans are believed to be better than models at generalization, few detailed comparisons have been carried out to date to rigorously quantify such putative discrepancies. Here we set out to document human ability to categorize objects in a challenging core object recognition task (DiCarlo, Zoccolan, & Rust, 2012) and to see how current state-of-the-art deep neural networks compare to humans on this task.

Methods

Image sets

In order to test generalization abilities extensively, we presented both humans and models with a broad range of stimuli, ranging from familiar naturalistic scenes to stimuli only familiar to humans but not models (e.g., artistic and blurry images) to stimuli that were unlikely to be part of their experience (e.g., block-shuffled and swirled images) (Figure 1). Our stimulus set spanned six styles: synthetic naturalistic images, referred to as ”HvM dataset” (Pinto, Cox, & DiCarlo, 2008), natural photographs from Microsoft COCO dataset (Lin et al., 2014), paintings, sketches, cartoons, and line drawings. HvM images were generated by randomly pasting a 3D object model onto a random naturalistic background (Pinto et al., 2008). COCO images were constrained to roughly match HvM images, namely: (i) only had one of the 10 tested object categories present, (ii) were matched in object size, and (iii) were cropped a square aspect ratio such that object placement was less biased to center. Artistic images were collected from free online resources.

Using HvM and COCO images, we also generated perturbed versions of the original images by dividing images into blocks and shuffling them around (3 sizes of blocks), blurring images (4 levels of blur), swirling images at the center of the target object (3 levels of swirl), and converting target object into its silhouette and further making a convex hull out of this silhouette. For HvM images, we additionally generated outlines of objects and their skeletonized representations, and also ”mosaic” images that were generated by pasting a 3D model on a light texture and applying neural style transfer with an image of a mosaic as a source style (Huang & Belongie, 2017).

Each set contained 12 images per each of 10 object categories (bear, elephant, person, car, dog, apple, plane, chair, zebra, bird).

Human testing

On each trial, the observer was presented (100 ms duration) an image containing a target object and asked to report its identity by choosing from two options that immediately appeared after the test image. Each participant completed 200 trials and were only allowed to participate once for a given set of stimuli (for instance, only once for HvM blur level 3 images). Approximately 10 responses per image were collected (i.e. 10 observers). All experiments were conducted on Amazon

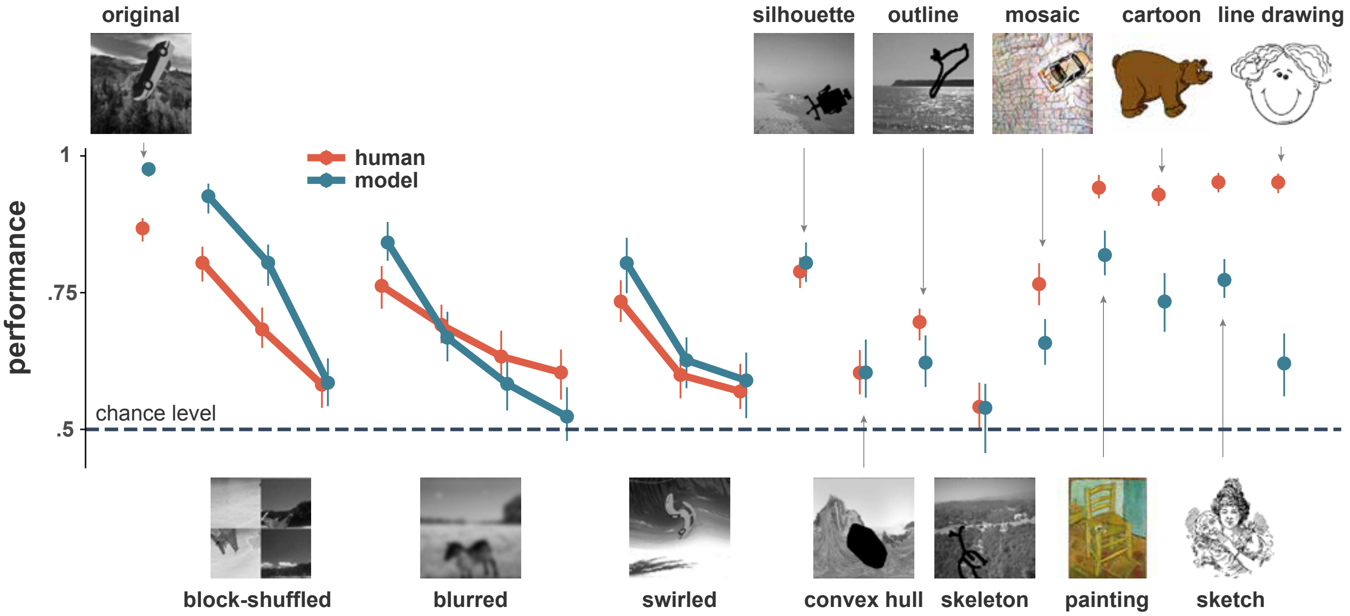


Figure 1: Models and humans perform similarly in visual domains that are either familiar (e.g., naturalistic stimuli) or unfamiliar to both (e.g., block-shuffled stimuli). However, humans are generally better in visual domains that could have been familiar only to them but not to models (e.g., cartoons). Note that models have been exposed only to ImageNet and original HvM images during training. Error bars depict 95% bootstrapped confidence intervals across images. (Responses to COCO-style images are not reported in this figure.)

Mechanical Turk platform and approved by Massachusetts Institute of Technology's Committee on the Use of Humans as Experimental Subjects (COUHES).

Model testing

We recorded artificial neural network (ANN) responses to every test image in the penultimate layer of ImageNet-trained ResNet-152 (He, Zhang, Ren, & Sun, 2016) and PNASNet (Liu et al., 2017) (state-of-the-art ImageNet model until very recently) using their TensorFlow Slim implementation. Since both ANN models lead to qualitatively very similar results, only ResNet-152 is reported. A 10-way logistic regression classifier was trained using these features in order to extract model's response probabilities for the 10 object categories used in the experiments. In our analyses, we contrasted training on only HvM or COCO images to training on images from all sets, thus the number of training images was matched between the two procedures. When comparing model performance on perturbations, 176 images per category were used to match the number of images in a 12-fold cross-validation procedure used on all images. When comparing model performance on style changes, 33 images per category were used (since there were fewer styles than perturbations). However, note that human-level performance can be achieved with much less training images; here we used the maximal possible number of train images to assess how well models can do in principle. Model response probabilities were further converted into two-way responses for each target-distractor pair to match the task format given to humans.

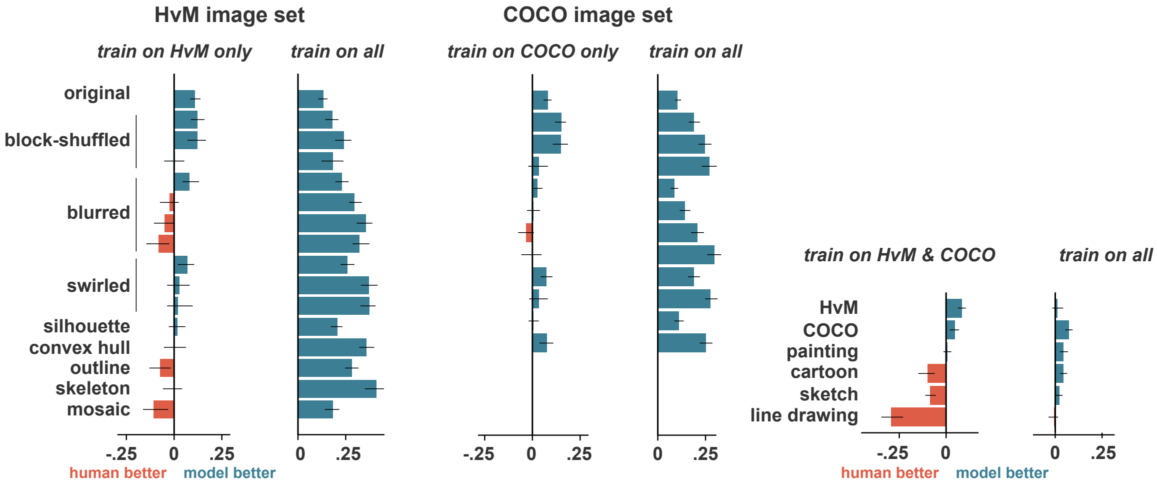
Results

Models generalize as well as humans on visual perturbations

We found that human performance across various visual domains was not uniform but rather spanned a broad range from nearly perfect to close-to-chance performance (Fig. 1, red dots). On the other hand, we observed that deep nets were less fragile than expected, generally closely following human response patterns, even without being trained on any of these perturbations (Fig. 1, blue dots).

We first explored these differences for visual perturbations (Figure 2a). We found that models matched human-level performance on most of these perturbations in the HvM-style image sets. Notably, humans were overall more robust to blur, in line with previous studies (Geirhos et al., 2017), and to mosaic and outline representations. We further asked whether models could have benefited from a rather restricted generative HvM space as it contained only 10 particular three-dimensional object models and, given sufficient training, a linear classifier could have capitalized on the idiosyncratic features of these precise models, resulting in an overestimation of deep nets' ability to generalize.

We therefore additionally tested humans and models on COCO images and their perturbed variants where each image contains virtually unique objects. Somewhat surprisingly, deep nets generalized as well or better on COCO-style images even though overall human performance did not degrade as much compared to perturbations on HvM-style images. We



(a) Image perturbations

(b) Image styles

Figure 2: Training models only on naturalistic images suffices to reach or exceed human-level performance across most image perturbations (a) but it does not generalize well to other images styles (b). However, object category information is linearly decodable from model features in all cases (”train on all” plots), suggesting that model representations are sufficiently rich to support broad generalization. Error bars depict 95% bootstrapped confidence intervals across images.

reasoned that both humans and models used contextual information in their judgments because, unlike in HvM-style images, image background was correlated with object category. Yet models apparently utilized on this contextual information more than humans, closing the human-model gap observed in HvM-style images.

Models generalize worse than humans across visual styles

Next, we compared human and model performance across image styles. In this case, model classifier was only trained on the original HvM and COCO images but not on any of the artistic images. We found that models were generally worse than humans in generalizing to unfamiliar visual styles (Fig. 2b). Curiously, models did not show an impairment to paintings, suggesting that, albeit non-photographic, they may share fairly similar statistics to natural photographs already familiar to models. Observe however, that this effect was only due to training on COCO images; training on HvM images alone retains human advantage (see Fig. 1).

Object representations are disentangled in model feature spaces

Since models appeared to be more sensitive to changes in visual style, we wondered if it reflected the lack of linearly separable representations in model feature space given the lack of exposure to such styles during model training. Alternatively, models might have learned sufficiently rich visual representations to support disentangled representations of object categories even when presented with unfamiliar visual styles, and those could be in principle utilized by a better decoder.

We therefore trained linear decoders using samples from all image sets. We found that a simple linear decoder could distinguish between all 10 object categories at or slightly above human level for all image styles simultaneously (Fig. 2, ”train on all” plots) while on perturbed images, model performance vastly exceeded human generalization abilities.

Discussion

Overall, we observed that both humans and these specific deep artificial neural networks are generally well-matched across a wide range of visual stimuli in a core object recognition task. Furthermore, we found that the deep nets contain sufficiently disentangled representations of object categories, even for image styles they have not encountered during their training. Our results therefore indicate that merely improving the decoder part of the model and retaining the ImageNet-trained encoder may be sufficient to achieve human-level generalization abilities out-of-the-box, at least on the range of visual image domains we have tested here.

On the other hand, while our stimuli covered multiple visual domains, our task remained fairly simple, where in each trial humans and models only had two choices and only ten object categories were present. More challenging versions of this task, such as multiple choice or free labeling, and more stringent evaluation metrics, such as behavioral consistency (Rajalingham et al., 2018), might reveal discrepancies between human and model ability to generalize that our metric might not have been sufficiently sensitive to.

Acknowledgments

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 705498 (J.K.), US National Eye Institute grants R01-EY014970 (J.J.D.), Office of Naval Research MURI-114407 (J.J.D), and grants from the Simons Foundation (SCGB [325500, 542965], J.J.D). We thank Rishi Rajalingham and Pouya Bashivan for constructive discussions.

References

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258.

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

Hochberg, J., & Brooks, V. (1962). Pictorial recognition as an unlearned ability: A study of one child's performance. *The American Journal of Psychology*, *75*(4), 624–628.

Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Iccv.*

Kornblith, S., Shlens, J., & Le, Q. V. (2018). Do better imagenet models transfer better? *arXiv preprint arXiv:1805.08974*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).

Liu, C., Zoph, B., Shlens, J., Hua, W., Li, L.-J., Fei-Fei, L., . . . Murphy, K. (2017). Progressive neural architecture search. *arXiv preprint arXiv:1712.00559*.

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS computational biology*, *4*(1), e27.

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