Feedforward deep neural networks diverge from humans and monkeys on core visual object recognition behavior

***Rishi Rajalingham (rishir@mit.edu)**, ***Elias B. Issa (issa@mit.edu)**, **Kailyn Schmidt (kailyn@mit.edu)**, **Kohitij Kar (kohitij@mit.edu)**, **James J. DiCarlo (dicarlo@mit.edu)**
Department of Brain and Cognitive Sciences
McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Abstract

A good model of human object recognition should mimic human behavioral responses at its output, including making the same pattern of errors over all images. We applied this straightforward visual Turing test to the leading feedforward computational models of human vision (hierarchical convolutional neural networks, HCNNs) and to a leading animal model (rhesus macaques) by comparing object identity reports for 240 images generated from 24 synthetic objects rendered with viewing parameter variation. Using high-throughput psychophysics in monkeys and humans, we tested all pairwise object discrimination tasks for each image. We observed that monkeys are highly consistent to humans in their image-level pattern of object confusions. Next, we found that all tested HCNNs were significantly less consistent with humans and with monkeys. This gap in consistency at the image level could not be rescued by primate-like retinal input sampling, choice of output decoders, or model training. Crucially, given that objects and images were in no way optimized to be adversarial to HCNNs, these results show that current HCNNs fail to replicate the image-level error patterns of primates. Going forward, high-resolution, image-level behavior could serve as a strong constraint for discovering models that more precisely capture the neural mechanisms of object recognition.

Keywords: visual object recognition; human; monkey; deep neural network;

Introduction

Specific models drawn from a large family of feedforward, hierarchical convolutional neural networks (HCNNs) can now match or exceed humans in absolute performance on object recognition tasks, and those models constitute a dramatic advance in our understanding of neuronal population response patterns at mid (V4) and high (IT) level of the ventral visual processing stream (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014). We previously observed that these particular HCNNs displayed similar patterns of object-by-object confusions as humans and monkeys (Rajalingham, Schmidt, & DiCarlo, 2015). However, this pattern of object-by-object confusion is not the most stringent behavioral test, as it does not capture the fact that some images of an object are more challenging than other images of the same object. To overcome this limitation, we here collected millions of behavioral trials to precisely measure object recognition error patterns for each image in humans, monkeys and models. We present this high-resolution behavioral metric as a stringent behavioral benchmark for models of human vision (a "visual Turing test").

Methods

Behavioral measurements

We characterized core object recognition behavior using a binary match-to-sample paradigm with 24 basic-level objects (see Figure 1a). In order to collect sufficient trials to reliably measure behavior on each image, we randomly selected a set of 240 images (10 images/object) to focus data collection on.

All human behavioral data were collected from human subjects on Amazon Mechanical Turk (MTurk) performing 276 interleaved, basic-level, invariant, core object recognition tasks. We pooled together trials from 1238 human subjects to characterize aggregate human behavior (pooled human). Four additional human subjects were held-out from this aggregation in order to create a held-out human sample. Analogously, four adult male rhesus macaque monkeys were tested on the exact same object recognition tasks, using a novel home-cage behavioral system (MonkeyTurk) which leveraged a web application running on a tablet. All procedures were performed in compliance with NIH guidelines and the standards of the MIT Committee on Animal Care.

We tested several publicly available trained HCNN models (AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), VGG (Simonyan & Zisserman, 2014), GoogLeNet (Szegedy et al., 2015), ResNet (He, Zhang, Ren, & Sun, 2016), and Inception-v3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016)). For each model architecture, features were extracted from the same images that were presented to humans and monkeys. We trained back-end classifiers on the machine feature representation, and varied the classifier type (SVM, MCC, or KNN), the number of subsampled features, and the number of training images.

Analysis

For each image i and each distracter object J, we computed an unbiased measure of object discrimination performance using an image-level sensitivity index $d'_{i,J} = Z(HR_{i,J}) - Z(FAR_{i,J})$, where $Z(.)$ is the inverse of the cumulative normal distribution, $HR_{i,J}$ is the the hit rate of image i presented against distracter J, and $FAR_{i,J}$ is the false alarm rate of images of class J. Since previous work has already examined the pattern of object confusions, we focused our analyses on

the object-independent variance of this metric by subtracting the corresponding object-level sensitivity index.

To measure the consistency of image-level behavioral patterns between any two visual systems x, y, we computed a noise-adjusted correlation, $\tilde{\rho}_{x,y} = \frac{\rho_{x,y}}{\sqrt{\rho_{x,x} \times \rho_{y,y}}}$ which normalizes the raw correlation between their behavioral patterns by the geometric mean of the split-half internal consistencies of each system (Pearson correlation between split halves of data).

Results

In this work, we measured the behavior patterns of humans, monkeys and HCNNs on a large set of binary object recognition tasks at the resolution of individual images. From these data, we asked which model systems accurately capture human behavior. As previously reported, all HCNNs (with backend classifier parameters chosen to maximize performance) are highly consistent with humans on an object-level behavioral metric (see Fig 1b). Figure 1c (left panel) shows the image-level behavioral consistency relative to the pooled human of all candidate models. The pooled monkey is highly consistent with the pooled human ($\tilde{\rho} = 0.702$), and almost as good as the held-out human pool on this image-level behavioral benchmark ($\tilde{\rho} = 0.765$). In contrast, all candidate HCNN architectures exhibit a significant gap in consistency to the pooled human on this metric (e.g. Inception-v3: $\tilde{\rho} = 0.383$).

Fig 1c (right panel) shows this behavioral consistency as a function of behavioral performance, for the pooled monkey and for all HCNNs; all tested choices of backend classifier parameters are shown for each of the HCNN architectures. Additionally, we modified the Inception-v3 model by 1) imposing primate-like foveal sampling of images at the model input (dark green), and 2) fine-tuning on naturalistic synthetic images (light green). While these modifications had expected effects on performance (e.g. fine-tuning increased performance), they did not improve behavioral consistency with respect to the pooled human. No tested instance of HCNN representation and decoder type was sufficient to capture human image-level behavior.

Discussion

We observed that macaque monkeys demonstrated highly similar image-by-image behavioral patterns as humans in the domain of core object recognition, further validating this animal model of vision. There was a small but significant difference in consistency between the monkey and an equal sample of humans which could hint at a true species difference. However, this gap could also be due in part to experimental differences between MTurk and MonkeyTurk. In contrast, we found that all tested state-of-the-art feedforward object recognition models could not replicate human behavior on this image-level visual Turing test. Although it had been known that these HCNNs models diverged from human behavior on specifically chosen adversarial images (Szegedy et al., 2013), a strength of our work is that we did not optimize images to induce failure, but instead randomly sampled the

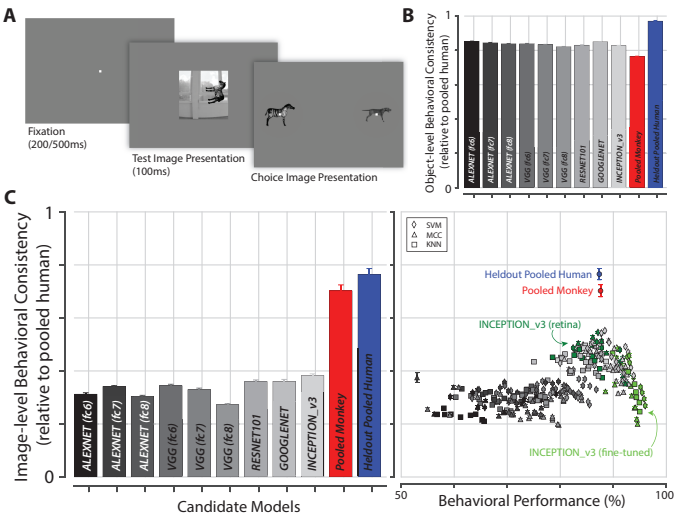

Figure 1: (a) Binary object recognition task for humans and monkeys. Object-level (b) and image-level (c, left panel) behavioral consistency for all candidate models. (c, right panel): Image-level behavioral consistency versus performance for all tested visual systems (each point is a model instance)

generative parameter space broadly. This suggests a generality of the finding that current feedforward HCNN models do not fully capture human core object recognition behavior, and high-resolution, image-level behavior could serve as a strong constraint for discovering models that more precisely capture the neural mechanisms underlying human object recognition.

References

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).

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput Biol*, *10*(11), e1003915.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Rajalingham, R., Schmidt, K., & DiCarlo, J. J. (2015). Comparison of object recognition behavior in human and monkey. *Journal of Neuroscience*, *35*(35), 12127–12136.

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

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).

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

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

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