Feed-forward deep neural networks diverge from primates on core visual object recognition behavior

Rishi Rajalingham*, Elias B. Issa[1]*, Kailyn Schmidt, Kohitij Kar, Pouya Bashivan, and James J. DiCarlo

Dept. of Brain and Cognitive Sciences, McGovern Institute for Brain Research, MIT
Department of Neuroscience, Zuckerman Mind Brain Behavior Institute, Columbia University[1]

Introduction

Humans can rapidly and accurately recognize objects in spite of variation in viewing parameters (e.g. position, pose, and size) and background conditions.

To understand this invariant object recognition ability, one can construct and test models that aim to accurately capture human behavior, including making the same patterns of errors over all images.

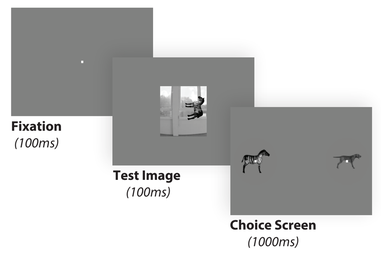
Here, we applied this straightforward visual "Turing test" to the leading feed-forward computational models of human vision (hierarchical convolutional neural networks, HCNNs) and to a leading animal model (rhesus macaques).

Our results reveal a failure of current HCNNs to fully replicate the image-level behavioral patterns of primates.

Methods

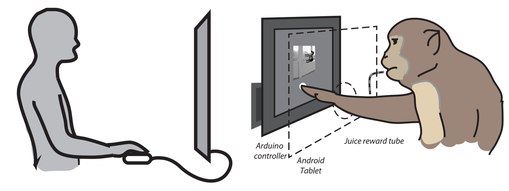
Behavioral paradigm:

We tested object recognition performance on a two alternative forced choice (2AFC) match-to-sample task using brief (100ms) presentations of naturalistic synthetic images of basic-level objects.



High-throughput psychophsyics:

Over a million behavioral trials across hundreds of humans and five monkeys were aggregated to characterize each species on each image. To collect such large behavioral datasets, we used Amazon Mechanical Turk for humans and a novel home-cage touchscreen system for monkeys.

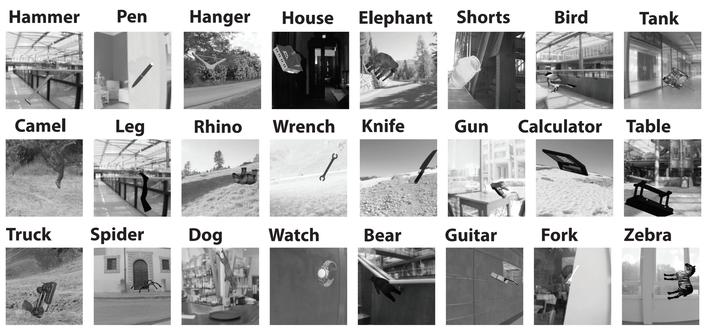


Amazon Mechanical Turk *"MonkeyTurk"*

Methods

Stimuli:

To enforce invariant object recognition behavior, we generated several thousand naturalistic synthetic images (2400 images, 24 objects, 100 images/object), each with one foreground object (by rendering a 3D model of each object with random viewing parameters) on a random natural image background.



Each object can produce myriad images under variations in viewing parameters, with some images being more challenging than others.

HARD *EASY*

Behavioral metrics:

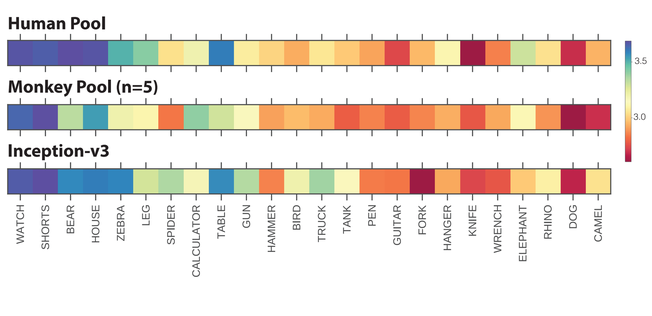
We characterized the core object recognition behavior of humans, macaques and HCNNs over a large set of images and distracter objects using several behavioral metrics. Each behavioral metric computes a pattern of unbiased behavioral performances, using a sensitivity index (d'), at the resolution of objects (O1, O2) and images (I1, I2). To isolate image-driven behavioral variance that is object independent, we normalized image-level behavioral metrics by subtracting the mean performance values over all images of the same object (I1c, I2c).

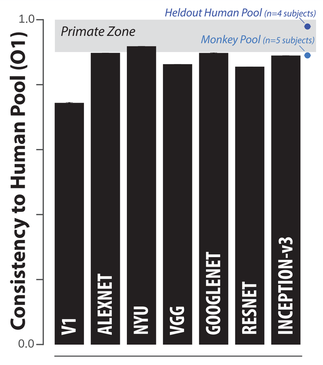
To measure the behavioral consistency between two systems with respect to a given behavioral metric, we computed a noise-adjusted correlation, which accounts for the reliability (i.e. internal consistency) of behavioral measurements.

$$\rho_{\text{normalized}} = \frac{\rho_{\text{explained}}}{\rho_{\text{explainable}}} = \frac{\rho_{x,\text{hum}}}{\sqrt{\rho_{x,x} \times \rho_{\text{hum,hum}}}}$$

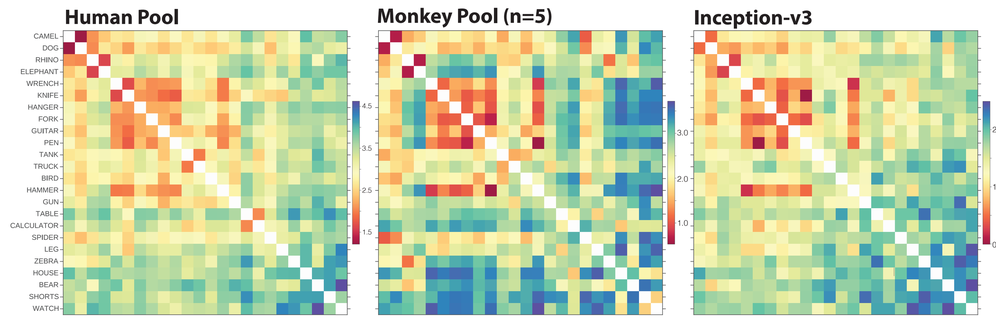
Results

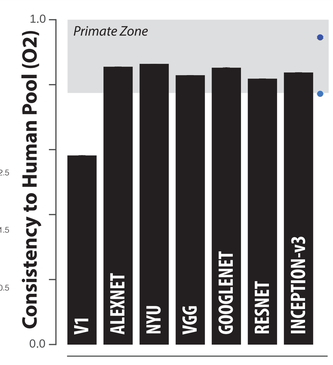
O1: Performance per object across distracters





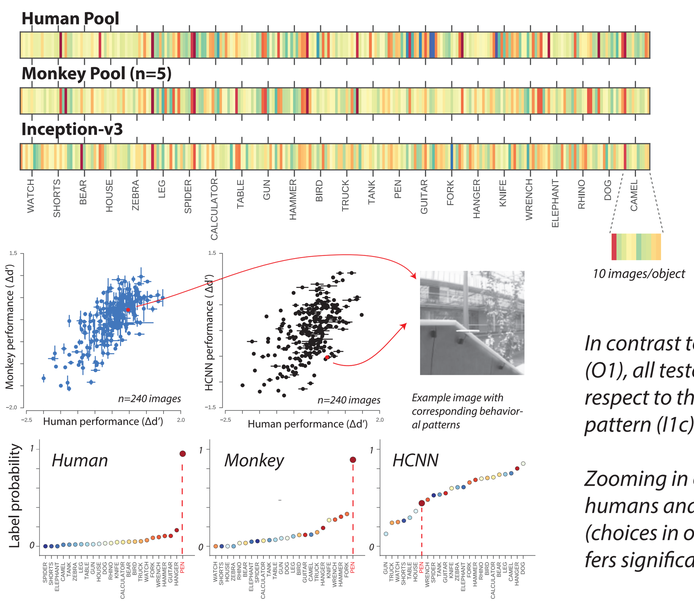
O2: Performance per object for each distracter (confusion matrix)

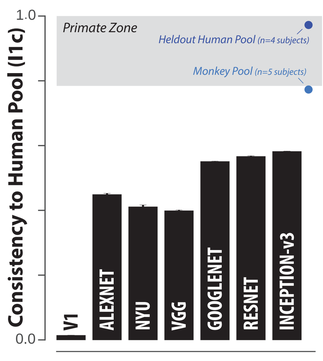




Consistent with previous work, macaque monkeys and all tested HCNN models are highly consistent with humans with respect to object-level metrics (O1, O2), while a lower level representation (V1) is less consistent.

I1c: Performance per image across distracters

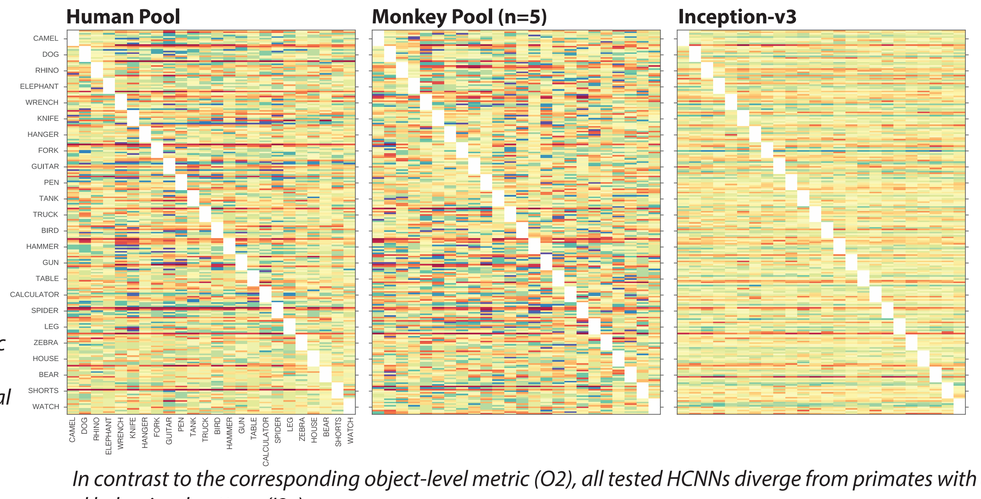


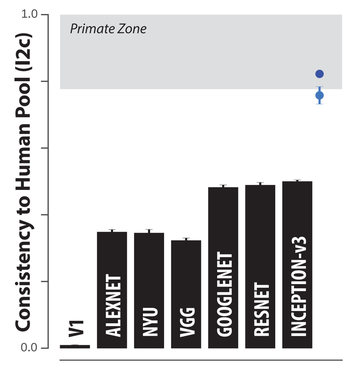


In contrast to the corresponding object-level metric (O1), all tested HCNNs diverge from primates with respect to the one-versus-all image-level behavioral pattern (I1c).

Zooming in on one example image of a pen, humans and monkeys make similar confusions (choices in object labels), while the HCNN model differs significantly.

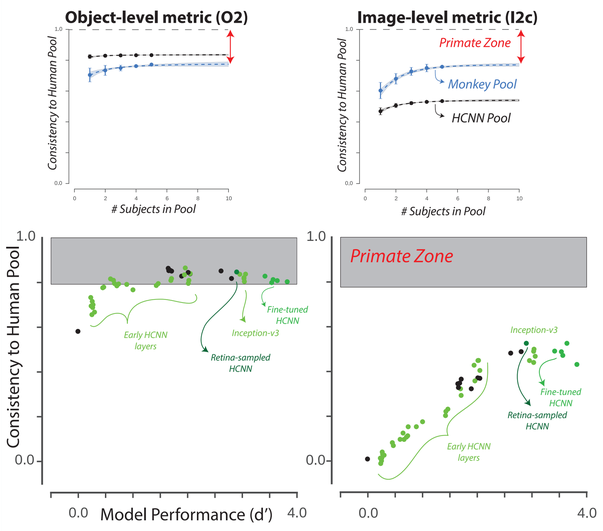
I2c: Performance per image for each distracter (confusion matrix)





In contrast to the corresponding object-level metric (O2), all tested HCNNs diverge from primates with respect to the one-versus-other image-level behavioral pattern (I2c).

All tested HCNN models fail to accurately capture the image-level behavioral patterns of primates.

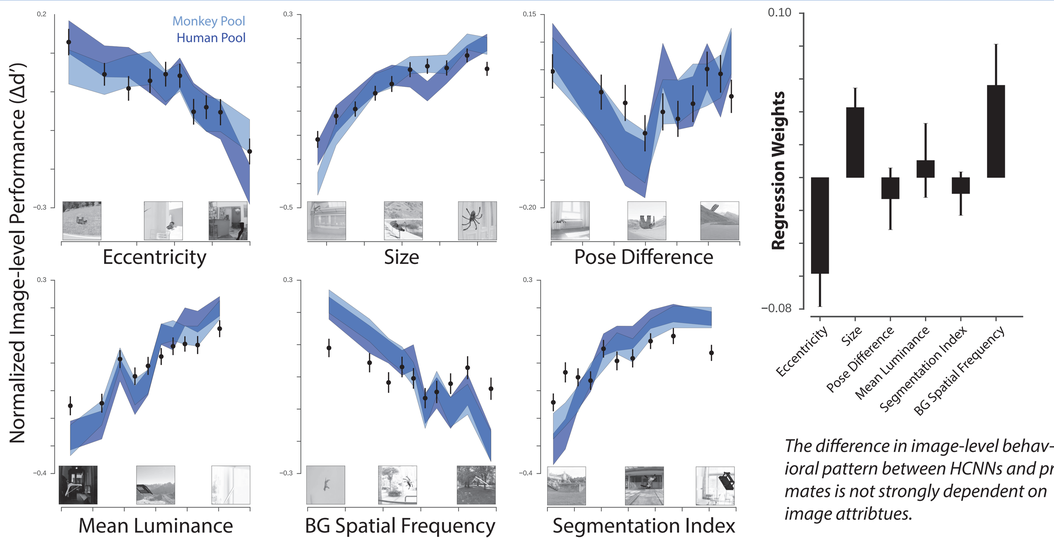


(Left, top row) For each behavioral metric, we defined a "primate zone" as the range of consistency values delimited by extrapolated estimates of consistency to the human pool of large (i.e. infinitely many subjects) pool of rhesus macaque monkeys and humans respectively.

Thus, the primate zone defines a range of consistency values that correspond to models that accurately capture the behavior of the human pool, at least as well as a monkey pool.

(Left) Consistency at the image level could not be easily rescued by pooling together several "model subjects," (top panels) *or simple modifications* (bottom panels), *such as primate-like retinal input sampling or additional model training on synthetic images generated in the same manner as our test set.*

(Right) Additionally, the gap in consistency could not be attributed to simple image attributes (e.g. viewpoint meta-parameters, low-level image characteristics.



The difference in image-level behavioral pattern between HCNNs and primates is not strongly dependent on image attribtues.

Conclusions

- In recent years, specific HCNN models, optimized to match human-level object recognition performance, have proven to be a dramatic advance in the state-of-the-art of artificial vision systems.

- HCNN models match primates in their core object recognition behavior with respect to object-level difficulty and object-level confusion patterns.

- HCNN models fail to match primate behavior in image-level difficulty and fail to match image-level confusion patterns.

- Attempts to rescue the image-level gap in consistency, including adding retinal sampling on the model front-end, adding various decoders on the model back-end, and supplementing model training with similar images, did not reduce the gap between models and primates.

- Furthermore, no particular property of the image (luminance, spatial frequencey) or of the object in the image (size, pose) was strongly correlated with model failure to capture primate image-level performance.

Taken together, these results suggest that high-resolution behaviour could serve as a strong constraint for discovering models that more precisely capture the neural mechanisms underlying primate object recognition.