Supplementary Material

A Data Collection and Processing

A.1 Neural Data Collection and Processing

A.1.1 Surgical implant of chronic micro-electrode arrays

In this study we have recorded chronic, in-vivo neural responses across both left and right hemispheres of 6 specific monkeys (3 used in model training and 3 used in model testing). First, we surgically implanted each monkey with a head post under aseptic conditions. After behavioral training, we recorded neural activity using multiple 10 × 10 micro-electrode arrays (Utah arrays; Blackrock Microsystems). A total of 96 electrodes were connected per array. Each electrode was 1.5 mm long and the distance between adjacent electrodes was 400 um. Before recording, we implanted each monkey multiple Utah arrays in the IT cortex. Array placement was guided by the sulcus pattern, which was visible during surgery. The electrodes were accessed through a percutaneous connector that allowed simultaneous recording from all 96 electrodes from each array. All behavioral training and testing was performed using standard operant conditioning (water reward), head stabilization, and real-time video eye tracking. All surgical and animal procedures were performed in accordance with National Institutes of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care.

A.1.2 Electrophysiological Recording

During each recording session, band-pass filtered (0.1 Hz to 10 kHz) neural activity was recorded continuously at a sampling rate of 20 kHz using Intan Recording Controller (Intan Technologies, LLC). The majority of the data presented here were based on multiunit activity. We detected the multiunit spikes after the raw data was collected. A multiunit spike event was defined as the threshold crossing when voltage (falling edge) deviated by less than three times the standard deviation of the raw voltage values. Our array placements allowed us to sample neural sites from different parts of IT, along the posterior to anterior axis. However, for all the analyses, we did not consider the specific spatial location of the site, and treated each site as a random sample from a pooled IT population.

Monkeys used for model training: We recorded neural data from three monkeys for training the CNNs. Chronic multi-electrode recordings from these monkeys' ventral visual cortex (recorded for various other research objectives) has been previously published (Majaj et al., 2015) (Hong et al., 2016) (Bashivan et al., 2019) (Rajalingham et al., 2020).

Monkeys used for model testing: We recorded neural data from three monkeys for testing the jointly optimized models. Chronic multi-electrode recordings from these monkeys' ventral visual cortex (recorded for various other research objectives) has been previously published (Kar et al., 2019) (Kar and DiCarlo, 2021) (Bashivan et al., 2019) (Rajalingham et al., 2020)

While majority of the data used in this study has been previously published to address different scientific questions (as mentioned above), we also collected recorded neural activity in one additional monkey (for model testing) exclusively for this study.

A.1.3 Neural recording quality metrics per site

Visual drive per neuron (d'_{visual}): We estimated the overall visual drive for each electrode. This metric was estimated by comparing the selected image responses of each site to a blank (gray screen) response.

$$d'_{visual} = \frac{avg(R_{img}) - avg(R_{gray})}{\sqrt{\frac{1}{2}(\sigma^2_{R_{img}} + \sigma^2_{R_{gray}})}} \tag{1}$$

Image rank-order response reliability per neural site (ρ^{IRO}_{site}): To estimate the reliability of the responses per site, we computed a spearman-brown corrected, split half (trial-based) correlation between the rank order of the image responses (all images).

Selectivity per neural site: For each site, we measured selectivity as the d' for separating that site's best (highest response-driving) stimulus from its worst (lowest response-driving) stimulus. d' was computed by comparing the response mean of the site over all trials on the best stimulus as compared to the response mean of the site over all trials on the worst stimulus, and normalized by the square-root of the mean of the variances of the sites on the two stimuli:

$$selectivity_i = \frac{mean(\vec{b_i}) - mean(\vec{w_i})}{\sqrt{\frac{var(\vec{b_i}) + var(\vec{w_i})}{2}}} \tag{2}$$

where $\vec{b_i}$ is the vector of responses of site i to its best stimulus over all trials and $\vec{w_i}$ is the vector of responses of site i to its worst stimulus. We computed this number in a cross-validated fashion, picking the best and worst stimulus on a subset of trials and then computing the selectivity measure on a separate set of trials, and averaging the selectivity value of 50 trial splits.

Inclusion criterion for neural sites: For our analyses, we only included the neural recording sites that had an overall significant visual drive (d'_{visual}), an image rank order response reliability (ρ^{IRO}_{site}) that was greater than 0.6 and a selectivity score that was greater than 1. Given that most of our neural metrics are corrected by the estimated noise at each neural site, the criterion for selection of neural sites is not that critical, and it was mostly done to reduce computation time by eliminating noisy recordings.

A.2 Human Data Collection

We measured human behavior (from 88 subjects) using the online Amazon MTurk platform which enables efficient collection of large-scale psychophysical data from crowd-sourced "human intelligence tasks" (HITs). The reliability of the online MTurk platform has been validated by comparing results obtained from online and in-lab psychophysical experiments (Majaj et al., 2015) (Rajalingham et al., 2015). Each trial started with a 100 ms presentation of the sample image (that contained a visual object embedded in a scene). This was followed by a blank gray screen for 100 ms; followed by a choice screen with the target and distractor objects (similar to (Rajalingham et al., 2018)). The subjects indicated their choice by touching the screen (for touch screen tablets) or clicking the mouse over the target object (for desktop computers). Each subjects saw an image only once. We collected the data such that, there were 80 unique subject responses per image, with varied distractor objects.

Human participants were compensated at a rate of approximately 4 USD per hour. The total amount spent was approximately 300 USD.

Human participants were greeted with the following messages upon clicking on the task on Amazon Mechanical Turk.

Welcome page instructions: "This HIT is part of a MIT scientific research project. Your decision to complete this HIT is voluntary. There is no way for us to identify you. The only information we will have, in addition to your responses, is the time at which you completed the survey. The results of the research may be presented at scientific meetings or published in scientific journals. Clicking on the 'SUBMIT' button on the bottom of this page indicates that you are at least 18 years of age and agree to complete this HIT voluntarily.

NOTE: Please close all other programs/taps while running this task to get the optimal system performance. Users on a suboptimal system can experience glitches that will lead to rejection. Also, low scores on this task will lead to rejection: make sure to read this instruction! Thank you for your interest! You are contributing to ongoing vision research at the Massachusetts Institute of Technology and the McGovern Institute for Brain Research. This task will require you to look at images on your computer screen and click to indicate a response for up to about 15 minutes, although we expect this to take about 5-10 minutes. If you cannot meet these requirements for any reason, or if doing so could cause discomfort or injury to you, do not accept this HIT. We encourage you to try a little bit of this HIT before accepting to ensure it is compatible with your system. If you think the task is working improperly, your computer may be incompatible. We recommend this task for those who are interested in contributing to scientific endeavors. Your answers will help MIT researchers better understand how the brain processes visual information."

Task instructions: "Please fixate at the center '+' sign which will appear once you press any key. Then, an image will be shown in the center of the screen. The image will contain an object. This

will be followed by a Choice Screen containing two objects. Please choose the option [Click on the object] that you think was present in the image"

B ADDITIONAL MODEL TRAINING AND TESTING DETAILS

In order to align model representations with primate IT representations while performing a classification task, we start by loading an ImageNet (Deng et al., 2009) pre-trained CORnet-S (Kubilius et al., 2019) as a base model. We add 8 new classification logits for the HVM image classes, and model hooks to extract representations from the first time step of the "IT" layer of CORnet-S. Next, we resume ImageNet training of the CORnet-S using stochastic gradient descent with a batch size of 128, a learning rate of 0.001, 0.9 momentum, and weight decay of 0.0001, but now for every gradient step we also include a batch of HVM images. For each HVM batch, a cross entropy loss is applied to the 8 new logits for classifying the HVM images, and a $log(1 - CKA(X, Y))$ loss is applied, where X is the activation matrix for the batch of representations extracted from CORnet-S "IT" layer and Y is the activation matrix of the corresponding image representations recorded from IT in the three training macaques. To generate a diversity of IT similarity in our models, we introduce random dropout of the CKA similarity loss which we call a neural loss ratio. A neural loss ratio of 1:1 corresponds to all HVM images having both a neural similarity loss and an HVM classification loss, while a neural loss ratio of 0.5:1 means that 50% of the time, the HVM batch only includes the HVM classification loss and no neural similarity loss. A neural loss ratio of 0:1 means that the model never has neural similarity gradients, only HVM classification gradients during the HVM batches. Finally, for experiments like 3 where we explore the effect of HVM label information on behavioral alignment, the HVM classification loss is set to 0 on all HVM batches. At no point during the training procedure does the model see COCO images or IT representations of COCO images.

In all reported experiments, model IT representational similarity training was performed on 2880 grey-scale naturalistic HVM image representations consisting of 188 active neural sites collated from the three training set macaques for 1200 epochs, or 26400 gradient steps on QUADRO RTX6000 GPUs, which took approximately 5 hours per training run. We introduce a small amount of data augmentation including the physical equivalent of 0.5 degrees of jitter the vertical and horizontal position of the image, 0.5 degrees of rotational jitter, and +/- 0.1 degrees of scaling jitter, selected to simulate natural viewing conditions. All models are in pytorch (Paszke et al., 2017), and training and testing is performed using pytorch lightning (Falcon et al., 2019).

Model IT representational similarity testing was performed on a total of three held out monkeys: Monkey 1 (280 neural sites) and monkey 2 (144 neural sites) on 320 held out HVM images with statistics similar to the training distribution, and monkey 1 (237 neural sites) and monkey 3 (106 active neural sites) on 200 full color natural COCO (Lin et al., 2014) images with different statistics than those used during training. IT neural similarity was computed as $CKA(X_{CORnet-S}, Y_{M_n})$ where $X_{CORnet-S}$ denotes the activation matrix of CORnet-S time step zero responses from the "IT" layer and Y_{M_n} is the activation matrix of test macaque n responses to all test images from an image set (HVM or COCO). The CKA score is computed for each test animal individually and normalized by the split half trial reliability of the animal's recordings, ie $CKA(X, Y)$ where X is the time and trial averaged recordings for 50% of the images presentations and Y is the time and trial averaged recordings for the other 50% of image presentations, bootstrapped over 10 random splitting seeds. Finally, the we average the individually computed and ceiled scores of each animal to give the Validated IT Neural Similarity scores reported in all experiments.

For performing white box adversarial attacks, we used untargeted projected gradient descent (PGD) (Madry et al., 2017) with L_∞ and L_2 norm constraints. Given an image x, this method uses the gradient of the loss to iteratively construct an adversarial image x_{adv} maximizing the model loss within an L_p bound around x. Formally, in the case of an L_∞ constraint, PGD iteratively computes x_{adv} as

$$x^{t+1} = Proj_{x+\mathcal{S}}(x^t + \alpha sgn(\nabla_{x^t} L(\theta, x^t, y)))$$

where x is the original image, and the $Proj$ operator ensures the final computed adversarial image x_{adv} is constrained to the space $x + \mathcal{S}$, here the L_∞ ball around x. Unless otherwise specified, we use $\|\delta\|_\infty = 1/1020$ constraints where $\delta = x - x_{adv}$ with 5 PGD iterations and a step size $\alpha = \|\delta\|_p / 4$ and report final top-1 accuracy. For more strenuous validation of our results we also compute full strength (ϵ) vs accuracy curves, and in that case we use 64 PGD steps instead of just five. The Adversarial Robustness Toolkit (Nicolae et al., 2018) was used for computing the attacks.

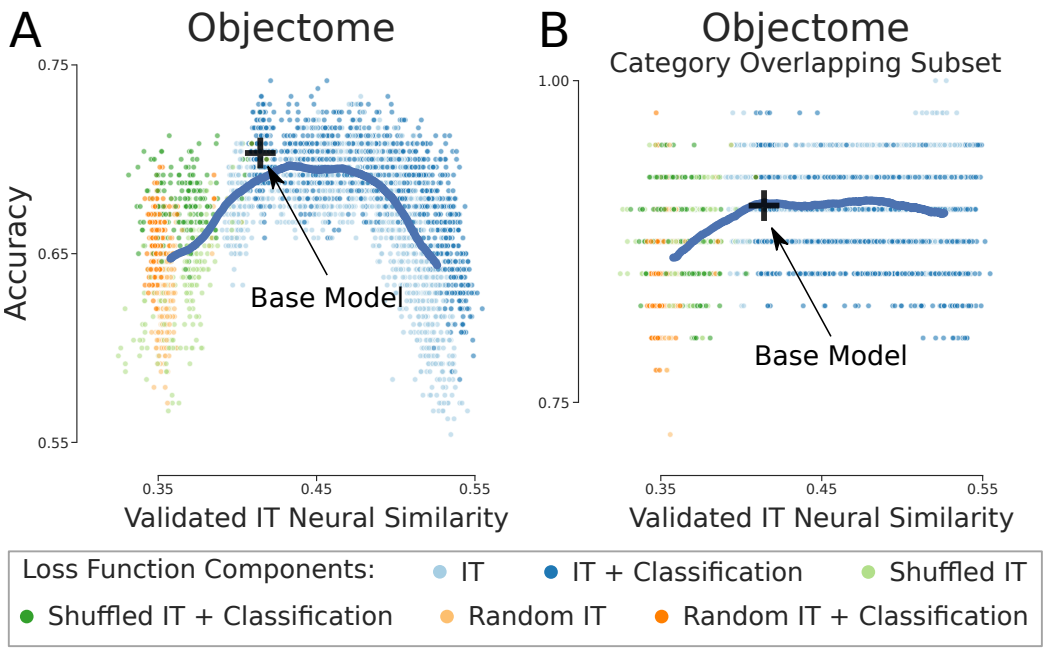

Figure A.1: **Higher levels of validated IT neural similarity are associated with decreasing classification accuracy on Objectome images over all, but not Objectome images with categories overlapping the IT training set categories. A**) Classification accuracy versus validated IT neural similarity for all model optimization conditions at all validation checkpoints reveals decreasing accuracy on the full Objectome image set which contains 20 categories not overlapping with the IT training set. **B**) Like in **A** but for the Objectome image set filtered to overlapping categories with the IT training set has relatively stable classification accuracy. In all plots, the black cross represents the average base model position, and the heavy blue line is a sliding X, Y average of all conditions merely to visually highlight trends. Five seeds for each condition are plotted.

C Additional Experimental Results

Here, to investigate why the behavioral alignment on Objectome Rajalingham et al. (2018) images decreases for higher levels of validated IT neural similarity, we use a linear probe to assess the (linear) class information content of Objectome image representations. Briefly, like for computing i2n behavioral alignment, feature representations are extracted for all image from the final average pooling layer of the network and partitioned into a train and test set. A linear classifier is fit to the train partition representations, and the classification error on the test set is reported. Results are shown in figure A.1. As can be seen, in the case of the full Objectome image set, classification accuracy decreases for higher values of IT neural similarity. Meanwhile, in the category overlapping subset, there is minimal loss in overall classification accuracy. Since i2n behavioral alignment is implicitly reliant on raw classification accuracy, this is at least one source of the decreasing behavioral alignment for Objectome images. Still, it is surprising to us that the network representations for categories beyond the IT fitting set are being degraded, and more research is needed to understand exactly what is happening to the underlying representations.

D Licensing Information

The MS COCO images dataset (Lin et al., 2014) is licensed under a Creative Commons Attribution 4.0 License. Adversarial Robustness Toolbox (Nicolae et al., 2018), Brain-Score (Schrimpf et al., 2018), Brain-Score associated datasets are under the MIT License. Pytorch (Paszke et al., 2017) is under a BSD license. Pytorch Lightning (Falcon et al., 2019) is under an Apache 2.0 license. CORnet-S (Kubilius et al., 2019) is under a GNU General Public License. ImageNet (Deng et al., 2009) is under a BSD 3-Clause License.