Adversarially trained neural representations may already be as robust as corresponding biological neural representations

Chong Guo [1] **Michael J. Lee** [1 2 3] **Guillaume Leclerc** [4] **Joel Dapello** [1 2 5] **Yug Rao** [6] **Aleksander Madry** [4 7]
James J. DiCarlo [1 2 3]

Abstract

Visual systems of primates are the gold standard of robust perception. There is thus a general belief that mimicking the neural representations that underlie those systems will yield artificial visual systems that are adversarially robust. In this work, we develop a method for performing adversarial visual attacks directly on primate brain activity. We then leverage this method to demonstrate that the above-mentioned belief might not be well founded. Specifically, we report that the biological neurons that make up visual systems of primates exhibit susceptibility to adversarial perturbations that is comparable in magnitude to existing (robustly trained) artificial neural networks.

1. Introduction

Deep neural networks (DNN) for computer vision are brittle in that their decisions are sensitive to small image perturbations which are targeted to modifying their outputs (adversarial attacks; (Szegedy et al., 2014; Goodfellow et al., 2015; Carlini & Wagner, 2017)). This is commonly regarded as an area for system improvement since similar brittleness has not been demonstrated in biological vision at comparable image perturbation strengths. Researchers have produced significant progress towards defense algorithms which improved the adversarial robustness of vanilla DNNs in standard computer vision tasks (Qin et al., 2019; Madry et al., 2018). Adversarial training in particular has been shown to both increase robustness on the original task (Croce & Hein, 2020) and to produce internal representations that better support transfer learning (Salman et al., 2020; Utrera et al., 2020) and image synthesis/manipulations (Engstrom et al., 2019b; Santurkar et al., 2019; Ledig et al., 2017). Yet, despite this progress, it is still widely believed that even the best of these networks are less robust than the 'gold standard' – the primate visual system. Is this assumption correct? In this study, we checked this assumption using primate neural recordings and ask: are the high level biological neural representations underlying primate object recognition truly more robust than existing artificial neural representations underlying current AI object recognition?

It is has not previously been possible to make robustness tests of primate visual neural networks comparable to those in artificial networks because the strongest adversarial attacks approximate the *worse-case* image perturbations by relying on detailed knowledge on each artificial network – knowledge that is still being developed for the biological system. Thus when that knowledge is limited, neuroscience experiments must rely on random sampling of image perturbation directions, which – given the high dimensionality of images – is unlikely to yield good estimates of adversarial (i.e. worst case) sensitivity within the time constraints of typical primate neuroscience experiments. Indeed, prior neuroscience work measuring the sensitivity of high-level visual representations in primates has focused on a restrictive set of image corruptions (i.e. scrambling) (Rust & DiCarlo, 2010) and quantifying invariance to transformations (position, size, context etc.)(DiCarlo et al., 2012; Logothetis et al., 1995; Ito et al., 1995; Tovee et al., 1994; Schwartz et al., 1983; Sáry et al., 1993; Ratan Murty & Arun, 2017). While informative, that work has not directly investigated the adversarial sensitivity of those neurons to small changes in pixel space.

Here we improved upon recent advances in mechanistic models of primate visual processing (Bashivan et al., 2019; Yamins et al., 2014; Dapello et al., 2020) to develop an experimental method to efficiently and iteratively measure the lower bound adversarial sensitivity of individual neural sites at the last stage of primate ventral visual processing pathway, the inferior temporal cortex (IT). Primate IT neurons are known to individually encode high-level features and to

[1]McGovern Institute for Brain Research, MIT [2]Department of Brain and Cognitive Sciences, MIT [3]Center for Brains, Minds and Machines, MIT [4]Computer Science and Artificial Intelligence Laboratory, MIT [5]School of Engineering and Applied Sciences, Harvard University [6]Purdue University [7]Department of Electrical Engineering and Computer Science, MIT. Correspondence to: Chong Guo <chongguo@mit.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

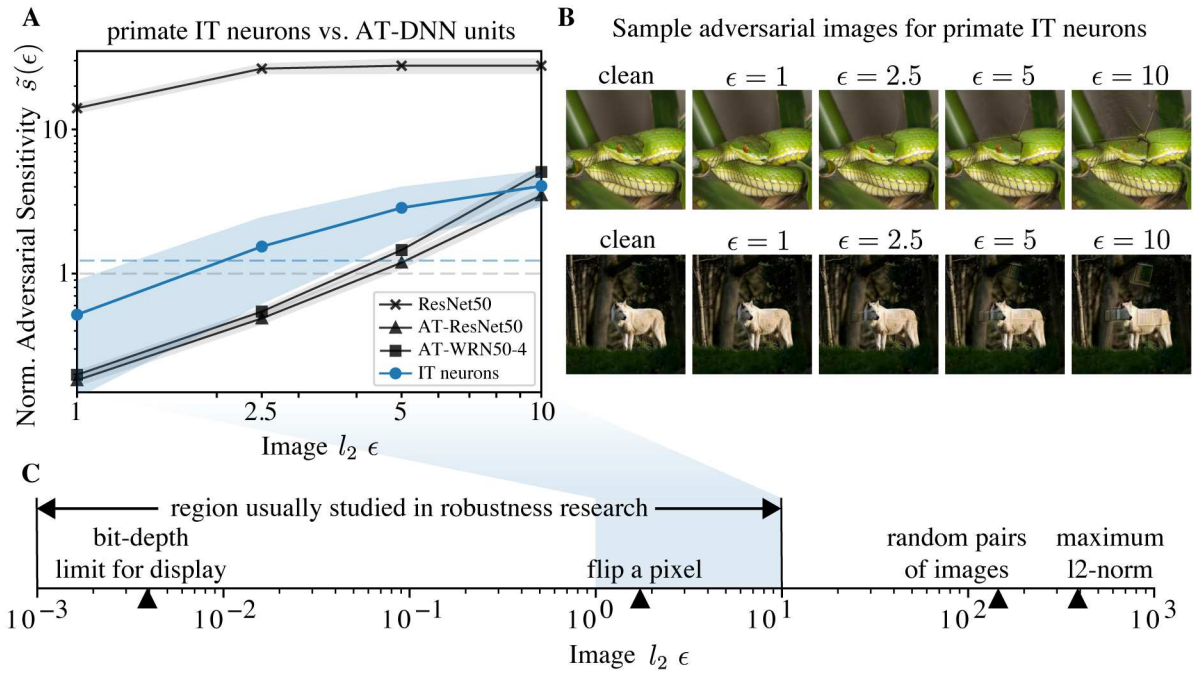

Figure 1. We measured the empirical lower-bound on adversarial sensitivity in biological neurons over a range of l_2-norm image perturbations relevant to current robustness research (blue section over the scalebar in **C**). **A.** Average normalized adversarial sensitivity $\tilde{s}(\epsilon)$ of IT neural sites (blue, mean±s.d.) is compared to that of features obtained from standard training on ResNet50, AT-ResNet50 and AT-WRN50-4 with $l_2\epsilon = 3$ (black, mean±s.d.). Standard deviation of either IT neural or DNN unit responses on clean images is shown (grey dashed line). The average absolute difference in IT neural responses evoked by a pair of random images is also shown for scale (blue dashed line). **B.** Adversarial images for two sample neural recording sites. **C.** The expanded scale for all feasible l_2-norm image perturbations for primate experiments, and notable perturbation sizes (dark triangles) are shown for reference on the horizontal-axis.

collectively underlie the perception and behavioral report of visual world latents such as object category and identity (DiCarlo et al., 2012; Logothetis et al., 1995; Tanaka, 1996; Miyashita, 1993; Majaj et al., 2015; Hong et al., 2016). For these reasons, IT can be loosely considered the functional equivalent of the layer just before the linear soft-max decoder in an artificial neural network.

For context, prior experimental work on IT neurons focused on much larger image perturbations that were motivated by the hypothesized computational goals of the primate ventral stream (e.g. estimate object category invariant of viewing conditions)(Logothetis et al., 1995; DiCarlo et al., 2012; Rolls, 2000). Based on such works, we and others in the field thought it unlikely that IT neurons would be sensitive to the much smaller, nearly human imperceptible adversarial perturbations used in current machine learning robustness research (**Figure 1**). This assumption has resulted in a gap in our knowledge of neural response properties within the local vicinity of any given image. Therefore in this work we attempted to bridge this gap by directly comparing the adversarial sensitivity of individual IT neural sites with individual units in state-of-the-art robust deep neural networks.

2. Result

Our primary goal was to measure the sensitivity of the response of individual IT sites to worst-case local pixel perturbations of visual stimuli. For each neural site i, we measure its response $r_i(x)$ to clean images $x \sim \mathcal{D}$, where $\mathcal{D}$ is the ImageNet training set (Deng et al., 2009). We define the i^{th} site's image-specific neuronal adversarial sensitivity as the maximal observed movement that $r_i(x)$ makes under an l_2-norm bounded image perturbation δ which is ϵ away from the original image x:

$$s_i(x, \epsilon) = \max_{||\delta|| < \epsilon} |r_i(x) - r_i(x + \delta)| \quad (1)$$

Marginalizing the image distribution D, we define the i^{th} neural site's absolute adversarial sensitivity as:

$$s_i(\epsilon) = E_{x \sim D}[s_i(x, \epsilon)] \quad (2)$$

To be able to compare individual IT neural sites with individual units in artificial neural networks, we use a dimensionless normalized adversarial sensitivity measure $\tilde{s}_i(\epsilon) = s_i(\epsilon)/\sigma_i$, where σ_i is the standard deviation of

the neural site's response over many different clean images (see **Method**). Conceptually, this standard deviation is the site's typical dynamic range and because the same response normalization is applied to each artificial neural site, this allows meaningful quantitative comparison between artificial network units (original units: arbitrary scalar response) and biological neural sites (original units: spikes per second). Lastly we note, while the adversarial sensitivity for units in artificial networks are measured on white-box attacks, the attacks for IT neural sites are found using an imperfect model of those sites. Because of this, the estimated adversarial sensitivity for any IT neural site is strictly a lower bound on its white-box adversarial sensitivity (see **Method**).

We now report the main result comparing the average normalized adversarial sensitivity for primate IT neural sites with the quantitatively comparable measurements of units in the corresponding feature layer of artificial neural networks (**Figure 1**). For clarity, we reserve the experimental details and the algorithm for empirically lower-bounding the adversarial sensitivity for primate IT neural sites in the method section (**Figure 4**).

For reference, we first show the sensitivity of artificial units from the last layer before the linear soft-max decoder in a standard ImageNet pre-trained ResNet50 (black crosses, **Figure 1A**). The classification performance of this network is highly sensitive to adversarial attacks (Engstrom et al., 2019a), and as expected, our measure shows that individual units from vanilla ResNet50 are highly susceptible to l_2 pixel perturbations: the magnitude of response perturbations are on average over 10-fold larger than each unit's baseline response variations across many clean images (grey dashed line). In comparison with units from ResNet50, we find that individual IT neural sites are approximately 10-fold less sensitive to the same magnitude of image perturbations (blue round markers, n=21, **Figure 1A**). So far this result is qualitatively consistent with the standard intuition that primate vision is more robust than standard DNNs (above).

But what about the comparison with adversarially trained (AT) DNNs? Specifically, we tested AT-ResNet50 and AT-WideResNet50-4 after ImageNet adversarial pre-training with $l_2\epsilon = 3$. Both networks have improved adversarial robustness on ImageNet and outperform the vanilla ResNet50 on a variety of transfer learning tasks (Salman et al., 2020). Consistent with this, individual units from both AT-ResNet50 and AT-WideResNet50-4 (black triangle and square, **Figure 1A**) are much less sensitive than units from vanilla ResNet50. Surprisingly however, when we compared to IT neural sites, we discovered that units from both robust DNNs are slightly *less* sensitive to adversarial perturbations than IT sites. This is all the more surprising as our measurements on the biological neurons is an experimental *lower bound* on their true adversarial sensitivity since our attacks are discovered using an approximate model of IT (see **Method** and **Figure 4**). On the other hand, the adversarial sensitivity for each unit in the DNNs is measured using white-box PGD-attacks by taking the maximum over 100 randomly initialized attacks with 250 attack steps each (see **Method**). In other words, further experiments on these IT sites can only lead to the same result – that current robust network units are already *at least as* robust as these biological network units.

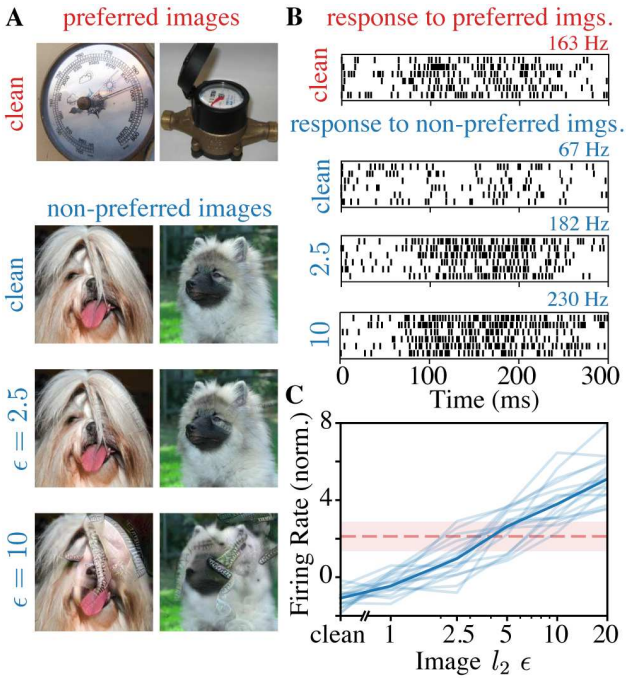

Figure 2. Category preference of individual IT sites can be adversarially attacked. **A.** Example clean images from preferred and non-preferred ImageNet categories for a representative IT neural site shown on the left, and the adversarially perturbed images at two ϵ values. **B.** Raw spike rasters associated with clean preferred images, clean non-preferred images and adversarially perturbed images at $\epsilon = 2.5$ and 10. **C.** Adversarial perturbations on non-preferred images (in blue) are able to drive firing rates past that of naturally occurring preferred images (red dashed line, $mean \pm s.d.$) at $\epsilon = 4.0 \pm 1.6$ (2.5 for the sample site), and by $\epsilon = 10$ turn non-preferred images into "super-stimuli" (individual sites in light blue, average in dark blue n=17).

The ability to adversarially attack biological neurons at such small perturbations levels can generate some highly counter-intuitive neurophysiological phenomena. Traditionally, IT neurons are known to demonstrate category/object selectivity (Gross et al., 1972). Neurons that respond to images of human faces, hands or specific animate or inanimate objects have been reported throughout IT cortex (Kanwisher et al., 1997; Tsao et al., 2003; Downing et al., 2001; Popivanov et al., 2014; Kornblith et al., 2013; Bao et al., 2020). Although the field no longer regards individual IT neurons as

pure object-category detectors, it is still thought that rank-ordered object-category preference is an important single unit property that underlie invariant object recognition (Li et al., 2009). Thus we ask the following simple question, how stable is "category preference" as a defining functional property of each IT neural site? We used each IT neural site's response over many clean images to identify its most and least preferred ImageNet categories (among 1000 categories), and difference in response between images of preferred vs. non-preferred categories. An example site is shown in (**Figure 2B**) along with two example images from its most and least preferred categories (**Figure 2A** top two rows), and the corresponding spike rasters show the clear difference in the density of spiking response following image presentation (preferred clean vs. non-preferred clean **Figure 2B** top two plots). Now we take arbitrary images from the non-preferred category (i.e. dog images for this example site) and we perform targeted adversarial perturbation on each of those images to change (here, increase) the neural site's response. We find that, by this reference calibration, perturbation slightly over $\epsilon = 2.5$, is on average sufficient to turn the site response to any non-preferred image into its same level of response to highly-preferred images (**Figure 2A,B** third rows). With a slightly larger perturbation $l2_2\epsilon = 10$, we found that we can drive the IT multi-unit site response to 310 ± 60Hz on average (n=17), effectively turning arbitrary non-preferred images into "super-stimuli" for these IT neural sites (**Figure 2A,B**, last rows). This exceeds the average response of these neurons to their 'most-preferred' image categories (216 ± 60 Hz). Interestingly, upon visual inspection, the super-stimulus do not usually even conform to the semantic categories from the clean preferred categories. This suggests that single-neuron's 'category/object selectivity' is not a locally stable functional property and highlights the insufficiency of concepts such as 'object detectors' and 'category selectivity' in building an accurate understanding of higher-level visual encoding.

It is surprising that primate IT neurons, which are approximately six anatomical stages of visual processing deep in the brain, were responsive to perturbations as small as $\epsilon = 1$, a barely noticeable change for humans (see **Appendix 2**) and smaller in magnitude than changing a single pixel from black to white ($\epsilon = 1.73$) (**Figure 1**). Are all IT neurons susceptible to adversarial attacks, or could the average results above be due to just a few strongly modulated neurons? We found that, while the sensitivity level varied, all recorded IT sites were similarly sensitive to adversarial attacks and each has a significantly positive slope between $\epsilon = 1$ to 10 (**Figure 3A**). We also analyzed the image-specific sensitivity, $s_i(x, \epsilon)$, by recording additional trials (10 repeats) to obtain a cleaner estimate on a subset of 100 images for three neural sites. We illustrate those results by showing sensitivity curves for each of 50 starting images for one typical IT neural site (**Figure 3B**). Here again, while the sensitivity curves for individual images varied, we found that most had significantly positive slopes measured between $\epsilon = 1$ to 10. Taken together, our results suggest that adversarial images can be readily found on all recorded IT sites and can be found very close to any clean image in the ambient image space (i.e. adversarial samples for biological neurons are dense in the image space similar to that of artificial neural networks (Szegedy et al., 2014)).

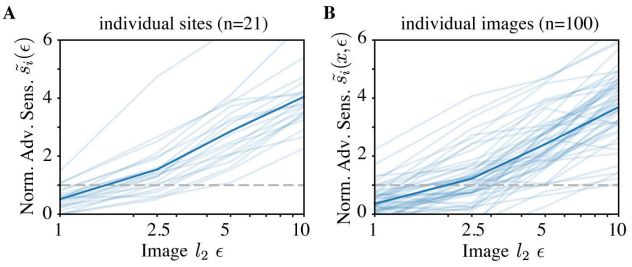

Figure 3. All tested individual IT neural sites are comparably sensitive to these image perturbations, and successfully perturbing images can be found near any starting (clean) image. **A.** Neural site-specific sensitivity curves (light blue) and the average over all sites, reproduced from figure 1 (dark blue). **B.** Image-specific sensitivity curves for a single example IT site. Each line (light blue) is the site's measured sensitivity to perturbations near a starting clean image (for visibility, 50 of 100 randomly selected starting clean images are shown). The dark blue line shows the average over all images tested for this site (i.e. one of of the light blue lines in A)

3. Method

3.1. Measuring adversarial sensitivity of IT neural sites

Measuring adversarial sensitivity $s_i(\epsilon)$ for each IT neural site i requires us to maximize the observed $|r_i(x) - r_i(x + \delta_i)|$ by finding better and better neuron-specific perturbation δ_i on every image x (**Figure 4**). To do so, we start with perturbations generated from a random baseline model of IT, and iteratively fit the model to observed IT attack responses until convergence. We performed extensive experiments to screen for the best baseline model of the IT sites (see **Appendix 1** for details on our screening procedure). We found the best baseline model is an adversarially pre-trained ImageNet model AT-ResNet50 ($l2\epsilon = 2$) that is linearly-mapped with channel-factorized weights from layer 4.0 to a 21 dimensional output layer to model the IT neural sites (**Figure 4 box 0**. The model parameters θ_t at $t = 0$ includes the pre-trained model parameters ω and the linear IT mapping weights ψ. Linear IT mapping weights ψ are factorized and initialized as channel weight $w_c \sim N(0, 1)$ and spatial weights $w_s \sim N(0, 1)$ (Klindt et al., 2017).

Using this randomly mapped baseline model, we optimize attack images independently for each model neuron using PGD with random starts, 100 steps and step size=$\epsilon/3$

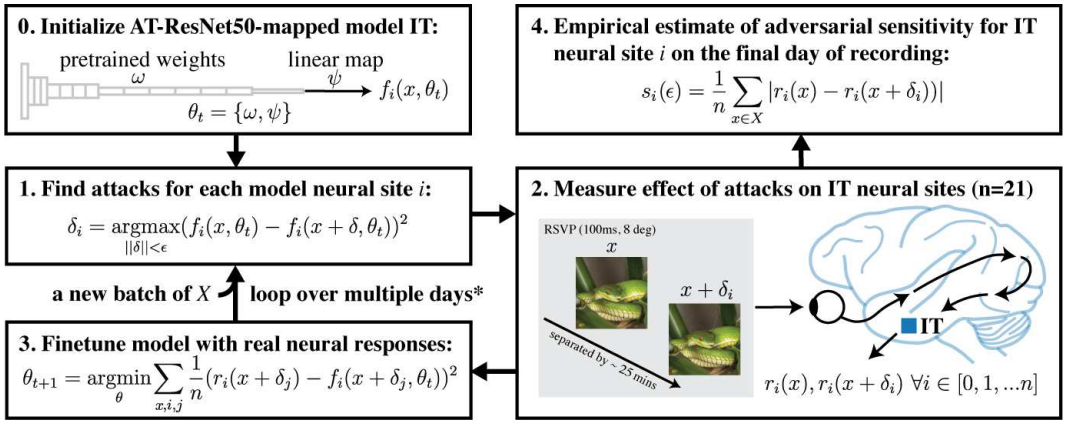

Figure 4. Adversarial training with primate IT neural recordings in the loop allows us to produce a robust functional copy of IT neural sites (box 0-3) and measure the lower-bound on each IT neural site's adversarial sensitivity (box 4)

(**Figure 4 box 1**). The clean images used on day 0, $X_{t=0}$, consist of 1000 clean images sampled from one of each of the 1000 ImageNet classes from the clean training set. Because for each neuron, the perturbation can either increase or decrease its firing rate, we perform PGD for both loss functions $\text{MSE}(r_i(x + \delta_i), 1000)$ or $\text{MSE}(r_i(x + \delta_i), -1000)$ and pick whichever one resulted in the largest predicted magnitude of neural perturbation.

After the attack images are found, we show both clean and attack images to a fixating monkey with two 99-channel Utah arrays (1.5mm, 400 pitch, Iridium Oxide coated electrodes) implanted in anterior and central IT (Blackrock Neurotech, Salt Lake City, UT). The visual stimuli are presented 8 degrees over the visual field for 100ms followed by a 100ms grey mask as in a standard rapid serial visual presentation (RSVP) task. The presentation order for all images (clean and perturbed) are shuffled across the experiment, which lasts for 5 hours on average. Given this design and the number of images, the average temporal separation between a clean image and its perturbed pair is 25 minutes. Thus image-specific response adaptation is unlikely to explain any aspects of these results. A minimum of two repetitions are shown for each image. For **Figure 1A**, we report IT sensitivity from the last day of experiment, which sampled 882 unique images per perturbation ϵ (i.e. 42 images per neural site). For each neural site i, we measure the total number of spikes between 70ms-170ms after image presentation for both the clean $r_i(x)$ and perturbed $r_i(x + \delta_i)$ trials (**Figure 4 box 2**).

At the end of the first recording session, we use both $(x, \vec{r}(x))$ and $(x + \delta_i, \vec{r}(x + \delta_i))$, $x \in X_{t=0}$, to train $\vec{f}(x, \theta_t)$ end-to-end with gradient descent and update θ_{t+1} (**Figure 4 box 3**). After this, a new batch of clean images is queued for the next day (X_{t+1}) and we repeat step 1 to 3 over a total of 6 days. As the model of IT $\vec{f}(x)$ improves, the adversarial attack δ_i solved for each neuron i should become better and better at generating larger perturbations in IT $|r_i(x) - r_i(x + \delta_i)|$ (**Figure 4 box 4**). Indeed, we tracked the measured neural perturbation magnitude and saw an consistent improvement over days (**Figure 5**). This suggests that the $l_2\epsilon = 1$ attacks from the model IT improved and transferred successfully onto real IT neurons. We note the perturbations achieved with our model is significantly larger than that achieved with a model-free method (Gaussian noise or linear interpolation to another image class), which confirms our intuition outlined in the introduction and

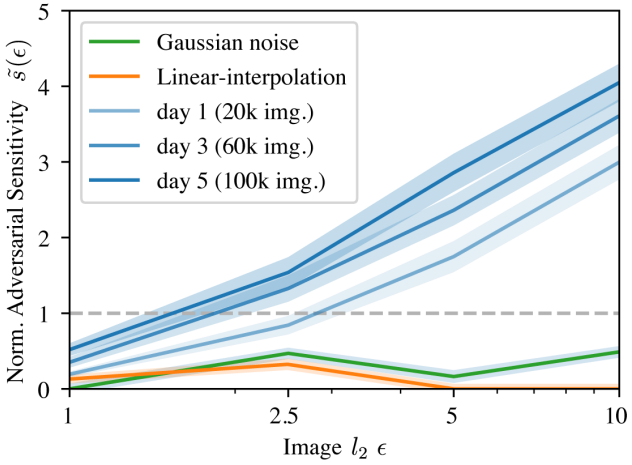

Figure 5. Normalized adversarial sensitivity lower bound improves over a number of days (blue) and is better than model-free perturbations such as Gaussian noise (green) or Linear-interpolation (orange). Sensitivity curve from day 0 where no mapping has occurred is not shown. Model-free measurements are measured as normalized root-mean-squared values which upper bounds the mean absolute differences.

explains why the field has systematically underestimated the sensitivity of neurons to local image perturbations. At the smallest perturbation $\epsilon = 1$, between the second (day 1: 20k samples for model tuning) and the sixth day of experiment (day 5: 100k samples for model tuning) there was a 2.7 fold increase in realized perturbation size. We report the normalized estimate $\tilde{s}_i(\epsilon)$ from the last day of experiment in all other figures as it is the highest lower bound we could obtain.

Lastly, in the case of biological neurons we use a negatively biased estimator for the average absolute neural response movement to avoid overstating the sensitivity of IT neurons. For x_j where $j = 0, 1, ..., M$, out of M number of images: $E_j[|r_i(x_j) - r_i(x_j + \delta_{i,j})|] = \frac{1}{M}\sum_j \text{sign}(r_i(x_j) - r_i(x_j + \delta_{i,j}))(r_i(x_j) - r_i(x_j + \delta_{i,j})) \geq \frac{1}{M}\sum_j \text{sign}(f_i(x_j, \theta) - f_i(x_j + \delta_{i,j}, \theta))(r_i(x_j) - r_i(x_j + \delta_{i,j}))$. In the presence of measurement noise, taking the average of the absolute value of response change will result in a positive bias. Therefore, the last equation is used for the estimation of average absolute neural response movement over images for each site. This estimator becomes unbiased if the model of IT site $f_i(x, \theta)$ from the *previous day* predicted all the directions of neural movement correctly.

3.2. Measuring adversarial sensitivity of individual artificial neurons

In the context of robust machine learning, evaluation of the sensitivity of a network is usually done by quantifying the accuracy of a model on adversarially perturbed images. Here we wish to compare the sensitivity of neurons in IT to sensitivity of units within an deep neural networks. The adversarial sensitivity of a single network unit i is defined as:

$$s_i(x, \epsilon) = \max_{||\delta|| < \epsilon} |h_i(x) - h_i(x + \delta)| \tag{3}$$

where x is a given image, h_i the activation of the i^{th} network unit of the penultimate layer (i.e. average pooling layer in ResNet50s). We use this layer in particular as this is the layer typically used for transfer learning and insofar contains a high-level representation of the pixel input learned by the network.

Besides taking the best out of 100 independent runs for solving the adversarial images for each unit, we also introduced multiple methods to drastically improve convergence beyond the basic PGD typically used for solving adversarial images:

- Optimizing $\max_{||\delta|| < \epsilon} h_i(x) - h_i(x + \delta)$ and $\max_{||\delta|| < \epsilon} h_i(x + \delta) - h_i(x)$ separately significantly reduces the chances to be stuck at saddle point. In most cases the latter one produces better solutions but we always attempt both.
- We observed that solving for larger ϵ converges faster. Therefore, for our evaluation we perform 250 steps of projected gradient descent first with a ball of radius 2ϵ and finally with one of radius ϵ. The relaxation of the first phase dramatically improves exploration of the search space and produces higher quality perturbations.
- For each optimization loop of 250 steps, we perform simulated annealing with restarts: we begin with steps of size ϵ and reduce them by 10% every time no progress is made. This schedule is repeated up to 4 times.

4. Related Works

Only a small number of studies have explored adversarial phenomena in human and non-human primate (Elsayed et al., 2018; Yuan et al., 2020; Zhou & Firestone, 2019). None of these directly measured adversarial sensitivity of neural representations in the regime studied by robustness research in the computer vision and machine learning communities ($l_2\epsilon < 10$). Yuan et al. attempted to adversarially attack neural and behavioral response in a two-way classification task in the primate using a vanilla ResNet101 (Yuan et al., 2020); however, the range of perturbations used in that study was between $l_2 = [21.4, 43]$, all of which are clearly visible to humans and substantially higher than what we explore here. Another study attempted to use an ensemble of non-robust networks to transfer attack time-limited human behavior on a three-way classification task using $l_{\text{inf}} = 32/255$ perturbation, which is equivalent to $l_2\epsilon = 48.7$ (Elsayed et al., 2018). As noted by Tramer et al., l_2 budgets greater than 19 fail to measure semantic similarity in images, and as they demonstrate, generating adversarial attacks to fool human behavior with an l_2 budget above 20 is trivially feasible by pasting in a target object without the aid of any behaviorial or neural models to guided the attack (Tramer et al., 2020).

This is also the first paper where a robust neural network is utilized for biological neural control under a limited budget (**Figure 2**). All existing works used vanilla networks, DNN or GAN, to synthesize images for single neuron or population control in V4 or IT (Bashivan et al., 2019; Ponce et al., 2019; Walker et al., 2019). The benefit of robust neural networks for biological neural control is that no additional image level prior needs to be manually enforced via additional loss functions. Relying on the network itself allows us to leverage additional data to discover the correct image level prior for neural/behavior control. This is possibly why we are able to control neurons with far less image perturbation budget, when all prior work used effectively unlimited budget for image manipulation.

5. Discussion

Here we provided the first experimental demonstration of adversarial sensitivity in biological neurons in a high-order brain area involved in visual object recognition. We find that the representations learned by adversarially trained artificial neural networks have already exceeded that of the corresponding biological neural representation in terms of their individual unit level adversarial robustness. Our results suggest that adversarial examples exist for all IT neural sites and that they are dense in the image space (i.e. nearby any starting image). Moreover, we demonstrated l2-norm perturbations as small as $\epsilon = 2.5$ could completely alter the category selectivity of recorded units, casting into doubt a traditional approach that vision scientists have relied on for decades for interpreting and cataloging functional neuronal types in IT.

This result confronts us with an apparent paradox: How is it that primate visual perception seems so robust yet its fundamental units of computation are far more sensitive than expected? One distinct possibility is that visual object recognition behavior in primate is actually not robust. This could be potentially explored with an iterative adversarial psychophysics experiment, similar to what we have done here for IT neurons. An alternative explanation is that there is an unknown error-correction mechanism at the population level in IT or in a down-stream area that decodes object identity. These hypotheses can be tested in subsequent experiments. We believe the current line of work could potentially lead to biologically-inspired solutions in ML robustness research, provide fundamental insights into the nature of adversarial phenomena in biological cognition, and perhaps provide new avenues to precisely modulate internal brain states without disrupting daily visual behavior.

References

Bao, P., She, L., McGill, M., and Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, July 2020.

Bashivan, P., Kar, K., and DiCarlo, J. J. Neural population control via deep image synthesis. *Science*, 364(6439), May 2019.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*, 2017.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In Iii, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 2020.

Dapello, Marques, Schrimpf, and others. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. *Adv. Eng. Educ.*, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*, 2009.

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

Downing, P. E., Jiang, Y., Shuman, M., and Kanwisher, N. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, September 2001.

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., and Sohl-Dickstein, J. Adversarial examples that fool both computer vision and Time-Limited humans. February 2018.

Engstrom, L., Ilyas, A., Salman, H., Santurkar, S., and Tsipras, D. Robustness (python library), 2019a. URL https://github.com/MadryLab/robustness.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Tran, B., and Madry, A. Adversarial robustness as a prior for learned representations. In *ArXiv preprint arXiv:1906.00945*, 2019b.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Gross, C. G., Rocha-Miranda, C. E., and Bender, D. B. Visual properties of neurons in inferotemporal cortex of the macaque. *J. Neurophysiol.*, 35(1):96–111, January 1972.

Hong, H., Yamins, D. L. K., Majaj, N. J., and DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.*, 19(4):613–622, April 2016.

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.*, 73(1):218–226, January 1995.

Kanwisher, N., McDermott, J., and Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.*, 17(11): 4302–4311, June 1997.

Klindt, D. A., Ecker, A. S., Euler, T., and Bethge, M. Neural system identification for large populations separating

what and where. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3509–3519. papers.nips.cc, 2017.

Kornblith, S., Cheng, X., Ohayon, S., and Tsao, D. Y. A network for scene processing in the macaque temporal lobe. *Neuron*, 79(4):766–781, August 2013.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. In *conference on computer vision and pattern recognition (CVPR)*, 2017.

Li, N., Cox, D. D., Zoccolan, D., and DiCarlo, J. J. What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? *J. Neurophysiol.*, 102(1):360–376, July 2009.

Logothetis, N. K., Pauls, J., and Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.*, 5(5):552–563, May 1995.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Majaj, N. J., Hong, H., Solomon, E. A., and DiCarlo, J. J. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *J. Neurosci.*, 35(39):13402–13418, September 2015.

Miyashita, Y. Inferior temporal cortex: where visual perception meets memory. *Annu. Rev. Neurosci.*, 16:245–263, 1993.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., and Livingstone, M. S. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4): 999–1009.e10, May 2019.

Popivanov, I. D., Jastorff, J., Vanduffel, W., and Vogels, R. Heterogeneous single-unit selectivity in an fMRI-defined body-selective patch. *J. Neurosci.*, 34(1):95–111, January 2014.

Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., De, S., Stanforth, R., and Kohli, P. Adversarial robustness through local linearization. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Ratan Murty, N. A. and Arun, S. P. A balanced comparison of object invariances in monkey IT neurons. *eNeuro*, 4 (2), March 2017.

Rolls, E. T. Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron*, 27(2):205–218, August 2000.

Rust, N. C. and DiCarlo, J. J. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci.*, 30(39): 12978–12995, September 2010.

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A. Do adversarially robust imagenet models transfer better? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Santurkar, S., Tsipras, D., Tran, B., Ilyas, A., Engstrom, L., and Madry, A. Image synthesis with a single (robust) classifier. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Sáry, G., Vogels, R., and Orban, G. A. Cue-invariant shape selectivity of macaque inferior temporal neurons. *Science*, 260(5110):995–997, May 1993.

Schwartz, E. L., Desimone, R., Albright, T. D., and Gross, C. G. Shape recognition and inferior temporal neurons. *Proc. Natl. Acad. Sci. U. S. A.*, 80(18):5776–5778, September 1983.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Tanaka, K. Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.*, 19:109–139, 1996.

Tovee, M. J., Rolls, E. T., and Azzopardi, P. Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque. *J. Neurophysiol.*, 72(3):1049–1060, September 1994.

Tramer, F., Behrmann, J., Carlini, N., Papernot, N., and Jacobsen, J.-H. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In Iii, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9561–9571. PMLR, 2020.

Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., and Tootell, R. B. H. Faces and objects in macaque cerebral cortex. *Nat. Neurosci.*, 6(9):989–995, September 2003.

Utrera, F., Kravitz, E., Erichson, N. B., Khanna, R., and Mahoney, M. W. Adversarially-trained deep nets transfer better. In *ArXiv preprint arXiv:2007.05869*, 2020.

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., and Tolias, A. S. Inception loops discover what excites neurons most using deep predictive models. *Nat. Neurosci.*, 22(12):2060–2065, December 2019.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.*, 111(23): 8619–8624, June 2014.

Yuan, L., Xiao, W., Kreiman, G., Tay, F. E. H., Feng, J., and others. Adversarial images for the primate brain. *arXiv preprint arXiv*, 2020.

Zhou, Z. and Firestone, C. Humans can decipher adversarial images. *Nat. Commun.*, 10(1):1–9, March 2019.

6. Appendix

6.1. Optimizing the adversarial base network for modeling IT

The training ϵ and layer for the base model of IT are chosen by three independent criteria: A. global representational similarity to IT as measured by CKA, B. cross-validated linear predictivity for IT responses, and C. how well does perturbations targeted toward a model layer transfers to IT neurons without any explicit mapping between the two systems. A and B are both performed on a separate set of neural recordings consisting of 12k images from the ImageNet training set. The best base models as measured by the three independent criteria are in complete agreement, which is layer 4.0 from $\epsilon = 2$ trained adversarial ResNet50. The resolution of detecting the best model appears to be better using the perturbation test of adversarial stimuli. Additionally we make the interesting observation that with AT-ResNet50s, adversarial training generally shifts the layer that is most similar to IT down to the deeper parts of the network with larger and larger training ϵ.

6.2. Perceptual discrimination of adversarial stimuli for IT neural sites

In a balanced experimental design where we ask subjects to detect the presence of adversarial perturbations designed for IT neural sites. We find that the accuracy of detecting a perturbation of $\epsilon = 1$ to be 51%, close to a chance level of 50% of random guessing in our design. Below we describe the detailed experimental protocol for this behavior result.

Briefly, Human subjects (n=50) were recruited on Amazon's Mechanical Turk Platform to conduct our behavioral experiments (n=13,8096 behavioral trials total) following COUHES guidelines. Subjects were free to do as many sessions as they liked (median number of trials per worker = 1,638 trials).

At the beginning of each session, subjects were given written instructions on how to successfully complete a single trial: first, two images were presented successively (100 msec durations each, with a 100 msec delay in between presentations in which a solid gray background was shown). Then subjects were instructed to report whether the two images were *completely* identical, or not (up to 10 seconds to respond). Images were presented on a neutral gray backgrounds, at approximately 8 degrees of visual field (based on assumptions of typical monitor sizes and viewing distances). Trials in which the subject failed to make a response within 10 seconds were discarded from our analysis.

On any given session (which consisted of 100 experimental trials), we balanced the number of positive trials and negative trials. Specifically, for any given perturbed image, we included 4 trials: perturbed-clean, clean-perturbed, clean-clean, and perturbed-perturbed. Such a scheme ensured that perturbed images showed up in all positions (first and second frame) and at the same rate as its original version, and that random guessing or choice biases would lead to an average accuracy of 50%. Thus, in each session, we were able to obtain an empirical estimate of the true detection rate and false detection rate with respect to 25 perturbed images.

By collecting data over many such sessions over many images, we were able to obtain subject-averaged estimates of detection rates (corrected for bias) for all of the perturbed images used in this experiment. We were then able to average these estimates by the perturbation ϵ values associated with each of the perturbed images.

We also included 'catch' trials in which two different base images were presented in sequence (leading to an 'obvious' choice of the two images being different), every 10 trials. By doing so, we could estimate a lower bound on the rate at which subjects were attending to the task (non-attending subjects would have 50% accuracy on catch trials) over the course of the session. On such catch trials, subjects had an average accuracy of 97.4% (0.95-CI: [97.1, 97.7]), suggesting the data we collected reflected subjects who understood the instructions and were highly attentive to the task (i.e. were not randomly guessing throughout the session).

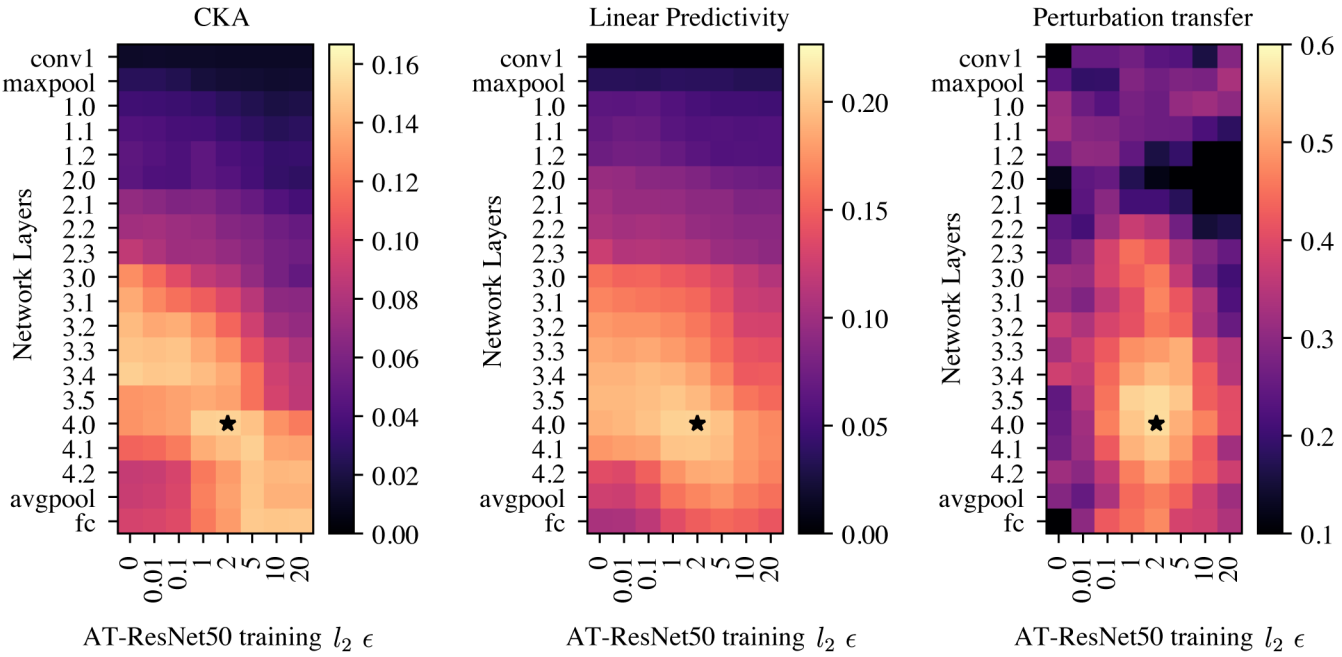

Figure 6. Selecting the layer (vertical-axis) and adversarial training ϵ (horizontal-axis) for the base mapping model of IT. Black asterisks denote the best layer/training ϵ combination selected using each of the three metric.