Primate Inferotemporal Cortex Neurons Generalize Better to Novel Image Distributions Than Analogous Deep Neural Networks Units

Anonymous Author(s)
Affiliation
Address
email

Abstract

Humans are successfully able to recognize objects in a variety of image distributions. Today's artificial neural networks (ANNs), on the other hand, struggle to recognize objects in many image domains, especially those different from the training distribution. It is currently unclear which parts of the ANNs could be improved in order to close this generalization gap. In this work, we used recordings from primate high-level visual cortex (IT) to isolate whether ANNs lag behind primate generalization capabilities because of their encoder (transformations up to the penultimate layer), or their decoder (linear transformation into class labels). Specifically, we fit a linear decoder on images from one domain and evaluate transfer performance on twelve held-out domains, comparing fitting on primate IT representations vs. representations in ANN penultimate layers. To fairly compare, we scale the number of each ANN's units so that its in-domain performance matches that of the sampled IT population (i.e. 71 IT neural sites, 73% binary-choice accuracy). We find that the sampled primate population achieves, on average, 68% performance on the held-out-domains. Comparably sampled populations from ANN model units generalize less well, maintaining on average 60%. These results are independent of the number of sampled units: models' out-of-domain accuracies consistently lag behind primate IT. These results suggest that making ANN model units more like primate IT will improve the generalization performance of ANNs.

1 Introduction

Ample behavioral evidence shows that humans are able to recognize objects in a wide variety of image distributions better than current computational models [1, 2, 3], indicating that yet-to-be-created ANN models (i.e. those more like the human brain) could generalize far better than current models. However it remains unclear which parts of the current ANN models are most in need of improvement, and what could be fixed to create those new models.

The standard scientific conceptual framework of how humans and non-human primates recognize objects is that, based on the retinal activity evoked by each image, the visual ventral stream produces a powerful feature set via hierarchical transformations from early visual cortex V1 through inferotemporal cortex (IT), and that features of the IT neural population are then linearly combined via downstream neural circuits to produce a class choice behavioral decision [4, 5, 6]. The processing from the image up to the IT population activity is referred to as the image "encoder" and the combination of IT features into the behavior is referred to as the "decoder". Accordingly, ANN models that are mapped onto this framework can be separated into encoder (features at their penultimate layer) and decoder (their linear transformation into class labels). These models assume the same

Submitted to 4th Workshop on Shared Visual Representations in Human and Machine Visual Intelligence (SVRHM) at NeurIPS 2022. Do not distribute.

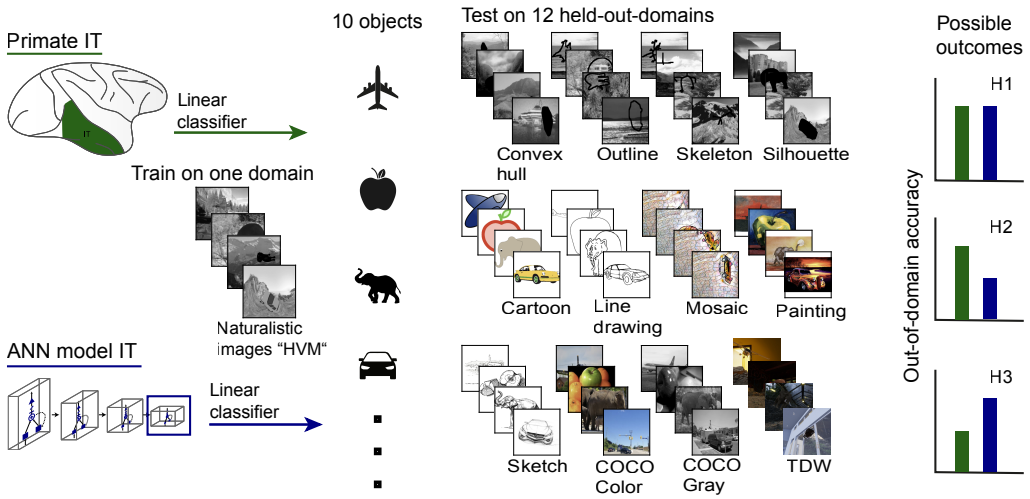

Figure 1: **Comparing the generalization performance of primate IT vs. ANN model representations with linear decoding.** We constructed image sets with 2400 images from 13 different domains, including e.g. naturalistic, cartoon, and sketch images. All domains contain objects from the same 10 categories, but in different image styles. To evaluate the goodness of representations (primate IT recordings, or ANN penultimate layers), we trained linear classifiers on the representations corresponding to one domain (naturalistic images, "HVM") and, without any further training, tested their ability to classify images from the 12 held-out domains into 10 categories. We can thus directly compare primate vs. model representations while keeping the readout fixed in order to determine whether model IT representations are as good as primate IT (H1), worse (H2), or better (H3).

kind of decoder as the standard neuroscience model (multi-class linear classifiers) and we will take advantage of that model-to-brain alignment assumption here.

In this work, we asked whether primate IT and model representations lead to different generalization performances when fitting a linear decoder on these representations from one image domain and then evaluating it in 12 unseen domains (Figure 1). Specifically, we use neural recordings from 71 primate IT sites in response to images from 13 different domains, and compare them to a "matched" number of penultimate units in several neural network models. We make 3 novel contributions:

1. We evaluate the generalization performance of a linear decoder fit on primate IT recordings (71 neurons) and find that it achieves 73% binary-choice accuracy in-domain (HVM) and 68% on average across 12 held-out domains.
2. We determine a model-specific scaling factor that lets us compare model-unit to primate-neuron, based on the number of model units necessary to reach the performance of a fixed number of primate neurons within an image domain.
3. Comparing the generalization performance of populations of primate IT neurons vs. a matched number of units in different models, we find primate IT to outperform current deep neural networks across different numbers of neurons, particularly for mosaic, outline, and silhouette image domains (average performance 60%).

2 Results

We first constructed an image dataset consisting of images categorized into 13 different domains: paintings, sketches, cartoons, line drawings, mosaics [1], natural photographs from Microsoft COCO [7] in color and grayscale, TDW [8], and synthetic naturalistic images ("HVM"; generated by randomly pasting 3D objects onto random naturalistic backgrounds [5]), and 4 domains obtained by converting HVM images: silhouette, convex-hull, skeleton, and outline [1] (see Figure 1 for examples). Each domain contains at least 60 images, with at least 6 images for each of the 10 object classes (apple, bear, bird, car, chair, dog, elephant, person, plane and zebra).

2.1 Primate IT neurons generalize to unseen image distributions

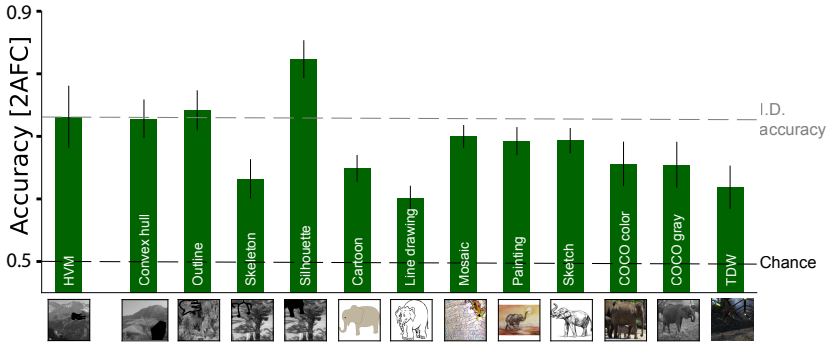

Figure 2: **Generalization performance of primate IT representations (71 sites) with a linear decoder.** We fit a linear classifier to primate IT representations (71 recording sites) corresponding to images from one domain (naturalistic images, HVM), and tested its performance on representations corresponding to held-out HVM images ("in-domain, I.D.") and images from 12 held-out domains ("out-of-domain"). Bars represent 2 alternative-forced-choice accuracy (chance = 50%) on each domain, and error bars are the standard deviation over train and test image sampling. Performance for each domain is well above chance, 68% on average over held-out domains (73% in-domain).

We implanted a macaque monkey three Utah micro-electrode arrays in inferior temporal cortex (IT; posterior, central, and anterior, $4.4mm \times 4.2mm$ span with 96 electrodes per array). 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. We pooled all IT sites and used only sites with consistent responses across image repetitions (Appendix A).

To evaluate the utility of IT representations for transferring to new domains, we fit a linear classifier to predict object labels. Specifically, we trained a ridge classifier on the IT population responses to 100 HVM images and evaluated it on 20 held-out images in-domain and 50 images for held-out-domains, using the class with the higher probability in each 2AFC trial (Appendix B).

Within-domain (HVM; testing on held-out images), this classifier achieves a binary-choice performance of 73% with 71 IT neurons (Figure 2). This result is roughly consistent with previous work which estimated that ~ 800 IT neurons are required to match human categorization performance [6]. Testing the same classifier (without further training) on images from the 12 held-out domains, the linearly decoded primate IT population on average maintains 68% accuracy and performs well above chance for all domains. The most difficult domains for the sampled IT neurons were line drawing (60%), TDW (62%), and skeleton (63%). These same domains had also been found to be behaviorally difficult for humans to recognize [1]. This is consistent with the standard neuroscience hypothesis that IT linear decodes underlie primate behavioral choice [5], and provided further confidence that our planned comparison of the IT population with various ANN neural populations would be meaningful.

2.2 ANN models generalize worse than primate IT

We next evaluated how the representations of current ANN models held up as a basis set for transfer performance compared to primate IT representations. We chose models based on the highest scores on Brain-Score [9, 10] IT benchmarks [5, 11, 12] and tested *CrossViT-18RA*, a dual-stream transformer with rotational invariance and adversarial training [13], *VOneResnet50*, where the first layer is a V1-like frontend model of the primate primary visual cortex [14], *resnet50-barlow*, a Resnet trained with self-supervised redundancy reduction [15], *CORnet-S*, a shallow recurrent network [16], and *alexnet* [17] as a standard baseline despite poorer Brain-Score performance.

To keep model units comparable to the limited number of primate IT recordings (71 sites), we estimated a scaling factor of how many units per model correspond to a primate neuron. To do this, we computed each model's in-domain performance (using the same linear classifier as for primate IT, Section 2.1) when sampling different numbers of neurons from the penultimate layer to determine the number of model units needed to achieve the identical performance of our primate IT population

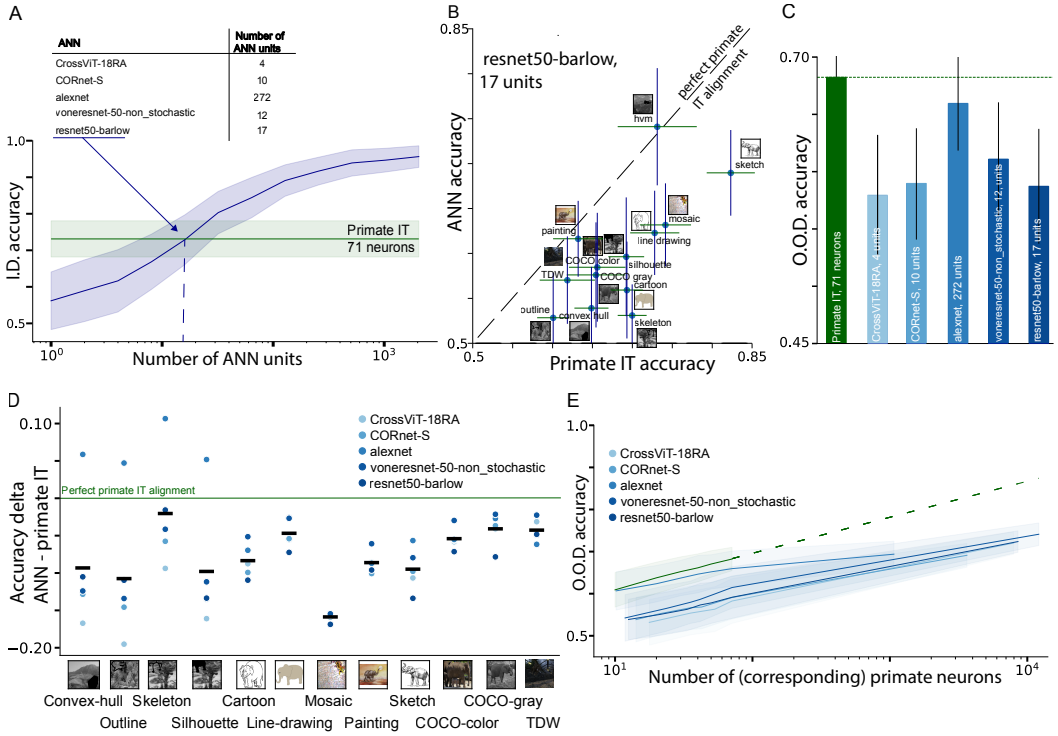

Figure 3: **ANN model unit populations do not generalize as well as comparable primate IT neural populations.** **A** Scaling model units to primate neurons: To keep the number of ANN model units comparable to the 71 primate IT sites, we determined how many units of each ANN are required to reach the same in-domain (I.D.) performance as with the primate IT representations. Sampling more units from the ANNs (x-axis, log scale) improves performance and eventually reaches the performance of 71 primate IT neurons. The table specifies for each ANN model how many units are necessary to reach this intercept. Plot is showing resnet50-barlow, see Figure B.1 for all models. **B** Alignment of ANN and primate IT generalization performance: We fit a linear decoder on the representations of 17 ANN units (sampled from the penultimate layer) in response to HVM images and evaluated its 2AFC accuracy on held-out HVM images and the 12 held-out domains. Primate IT accuracy as in Figure 2. Error bars are standard deviation over train and test image samples (and unit samples for the ANN). Showing resnet50-barlow, see Figure B.2 for all models. **C** Out-of-domain (O.O.D.) accuracy of primate IT representations compared to corresponding ANN unit representations: Averaging over the performances in 12 held-out domains (see *B*), we computed mean O.O.D. performances for each model. Error bars are standard deviation over domains. **D** Performance difference between ANN models and primate IT per domain: For each ANN and each domain, we subtract the performance of primate IT from ANN accuracy; i.e. deltas below zero mean that the ANN performs worse than primate IT. Horizontal green line is 0, i.e. no accuracy difference. **E** ANN models lag behind primate IT even when using more neurons: We plot the averaged out-of-domain accuracy of each ANN when using up to all the units in the penultimate layers (x-axis, log scale, number of units scaled to corresponding primate neurons following *A*). For primate IT, we logarithmically extrapolate up to 10.000 neurons. Shaded error bars are standard deviation over the 12 held-out domains.

(i.e. 71 recording IT sites). We found that the corresponding number of units varies between models, e.g. 17 in resnet50-barlow and 4 in CrossViT-18RA (Figures Figure 3A and B.1). Alexnet required an order of magnitude more units than other models to reach 73% in-domain performance (272 units). For the following analyses, we randomly sampled units from each model up to the number determined above.

Next, we evaluated the capabilities of each of these unit-matched model sub-populations in the same way we had evaluated the IT population (above): using a linear classifier trained on in-domain images

and then tested on each set of out-of-domains images. We found that models did not generalize as well as primate IT, with 60% average 2AFC accuracy across models (Figure 3B). While the models perform only 4 percent points worse on some domains (e.g. skeleton and COCO gray), mosaic, outline, sketch, and silhouette are the most difficult domains compared to primate IT (Figure 3C; p<10e-4, one-way ANOVA testing if mean differences of held-out-domains are different from one another). These results thus support H2 from Figure 1: ANN model representations do not generalize as well as comparable primate IT representations.

2.3 Performance gap between primate IT and ANNs is consistent when varying numbers of neurons

Since our analyses so far all operate on 71 primate IT neurons (and the corresponding number of model units), we tested how well the models would compare when varying the number of neurons in the two sub-populations (i.e. IT vs. ANN. While fewer IT neurons decreases accuracy on held-out domains, the accuracy of models with a corresponding number of units (determined by the scaling factor determined in Section 2.2) equally decreases, such that the gap between primate IT and models is consistent (Figure 3E).

In an exploratory analysis, we further compared the full ANN models (i.e. using all units in the penultimate layer) to the performance of larger primate IT neural populations (Figure 3E green dashed line). Even using all units of each model's penultimate layer leads to generalization accuracies below the primate IT performance (extrapolated via logarithmic $y = a \times log(x) + b$). For instance, VOneResnet-50 with all 2048 units achieves 75% average accuracy on the 12 held-out domains whereas we extrapolate primate IT with the corresponding number of 12,117 neurons to achieve 87%.

3 Discussion

Combining primate neural recordings, state-of-the-art artificial neural network (ANN) models, and linear decoding, we here found that primate IT neural populations generalize significantly better than corresponding ANN model unit populations. While linear classification from 71 primate IT neural sites maintains 93% accuracy on held-out domains relative to in-domain performance on average, ANNs maintain only 82% relative performance on average across models. ANNs fall most behind primate IT for mosaic, outline, and silhouette images. These results argue that making ANN models at their "IT" level more like primate IT should help close the behavioral generalization gap of ANNs observed in previous literature [1, 2, 3].

Data limitations. Our analyses were limited by a neural dataset of 71 primate IT recording sites, and a relatively small image count such that we were only able to train on 100 images. Since the number of training images was consistent across all systems under study (primate IT and ANNs), we do not expect this factor to influence relative performance differences. Dealing with the inherent sampling limitation for primate recordings, we estimated how many ANN model units correspond to the 71 recording sites, but it is possible that these scaling factors do not perfectly reflect the unit-to-neuron correspondence. We thus also compared the full number of model units to an extrapolated number of primate IT sites, which showed a consistent gap between primate and ANN performance. More neural recordings are necessary to validate extrapolation predictions.

Linear decoding. Following previous linking assumptions [5, 6], we used a linear classifier to transform neural representations into classification choices. Since we used the same linear decoding for primate IT and ANN populations, we can directly compare representations and find ANNs to lag behind primate IT generalization performance. This result suggests the "encoder" part of ANNs (transformations up to penultimate layer) as a culprit of non-human-like generalization performance but does not absolve the involvement of the linear decoder, i.e. the decoder might still require fixing. A strong test would be to evaluate whether primate IT populations with a linear decoder perfectly generalize to new domains and fully predict human behavioral choices in all situations.

Making models more primate IT-like. Our results argue that making ANN representations more like primate IT will improve generalization performance. How can we align ANN representations to inferotemporal cortex? Further improvements on ImageNet performance [18] no longer seem to improve neural alignment [9]. An alternative might be to optimize for neural recordings directly [e.g. 19, 20]. We predict such improvements on aligning models to the brain to yield concrete payoffs in transfer benchmarks, highlighting a synergistic link between brain science and machine learning with respect to generalization performance.

References

[1] Jonas Kubilius et al. "Can Deep Neural Networks Rival Human Ability to Generalize in Core Object Recognition?" In: *Cognitive Computational Neuroscience (CCN)*. 2018.

[2] Robert Geirhos et al. "Generalisation in humans and deep neural networks". In: *Neural Information Processing Systems (NeurIPS)*. 2018, pp. 7538–7550.

[3] Robert Geirhos et al. "Partial success in closing the gap between human and machine vision". In: *Neural Information Processing Systems (NeurIPS)* (2021).

[4] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. "How Does the Brain Solve Visual Object Recognition?" In: *Neuron* 73.3 (2012), pp. 415–434.

[5] Najib J Majaj et al. "Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance". In: *Journal of Neuroscience* 35.39 (2015), pp. 13402–13418.

[6] Ha Hong et al. "Explicit information for category-orthogonal object properties increases along the ventral stream". In: *Nature Neuroscience* 19.4 (2016), pp. 613–622.

[7] Tsung-Yi Lin et al. "Microsoft COCO: Common Objects in Context". In: *arXiv preprint* (2014).

[8] Chuang Gan et al. "ThreeDWorld: A Platform for Interactive Multi-Modal Physical Simulation". In: *Neural Information Processing Systems (NeurIPS)*. 2020.

[9] Martin Schrimpf et al. "Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?" In: *bioRxiv* (2018).

[10] Martin Schrimpf et al. "Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence". In: *Neuron* (2020).

[11] Kohitij Kar et al. "Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior". In: *Nature Neuroscience* (2019).

[12] Sachi Sanghavi and James J. DiCarlo. *Sanghavi2020*. Tech. rep. 2020.

[13] William Berrios and Arturo Deza. "Joint rotational invariance and adversarial training of a dual-stream Transformer yields state of the art Brain-Score for Area V4". In: *Cosyne Brain-Score Workshop*. 2022.

[14] Joel Dapello et al. "Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations". In: *Neural Information Processing Systems (NeurIPS)*. 2020.

[15] Jure Zbontar et al. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction". In: *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 12310–12320.

[16] Jonas Kubilius et al. "Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs". In: *Neural Information Processing Systems (NeurIPS)*. 2019, pp. 12785–12796.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Neural Information Processing Systems (NIPS)* (2012).

[18] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[19] Shahd Safarani et al. "Towards robust vision by multi-task learning on monkey visual cortex". In: *arXiv preprint* (2021).

[20] Joel Dapello et al. "Aligning Model and Macaque Inferior Temporal Cortex Representations Improves Model-to-Human Behavioral Alignment and Adversarial Robustness". In: *bioRxiv preprint* (2022).

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We are preparing a code and data release for a full publication
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplementary Material

A Primate recordings

Passive fixation task: We presented images on an LCD monitor (1920×1080 pixels, 60 Hz refresh rate; ASUS VS248H-P) placed in front of the animal. The images were presented on a grey background, in the central 8° of the animal's visual field. Standard rapid serial visual presentation (RSVP) protocol was used, with 100 ms on and 100 ms off durations, a 300 ms pretrial fixation period, and 6–10 images per trial. The animal was head-restrained, and trained to passively fixate on a central white dot (0.2°). Real-time eye-tracking (SR Research, EyeLink II) was employed to ensure that eye jitter did not exceed $\pm 2^\circ$, otherwise the trial was aborted and data discarded. For successful trials, the animal was rewarded with juice. Each image was presented pseudo- randomly once in an experimental block. The blocks were repeated multiple times, resulting in 44–61 repeat recordings of each testing image. Stimulus display and reward control were managed using the MWorks Software (http://mworks-project.org).

Electrophysiological recodings: During each recording session, neural activity was amplified ($1\times$ gain) and digitized (sampled at 20 kHz) using the the Intan RHD Recording Controller (Intan Technologies, LLC). The raw voltage signals were bandpass filtered offline using a second-order Elliptic filter (300 Hz to 6 kHz, 0.1 dB passband ripple, 50 dB stopband attenuation), before being thresholded to obtain the muti-unit spike counts. The threshold was set for each recording session and each recording site individually, at 3 standard deviations above the noise. To ensure accurate stimulus locking, the spikes were aligned to a photodiode trigger (Thorlabs, DET36A2) attached to the display screen.

Only IT sites that were deemed "stable" across recording days were retained. To pick these sites, we showed each image in multiple repetitions, and computed the similarity of one half of image repetitions (averaged) to the other half of repetitions of the same images (also averaged). In detail, we measured the internal consistency, defined as the Spearman-Brown corrected split-half reliability, of responses for a fixed set of 25 out-of-set naturalistic images (termed the normalizer set) shown at the beginning of every recording session ($r > 0.75$), and for HvM images shown in a separate experiment ($r > 0.7$). Of 288 recording sites (three arrays $\times$ 96 electrodes), we retained 71 sites across anterior, central, and posterior inferior temporal cortex.

The multi-unit response to an image was taken as the average spiking rate in the 70–170ms temporal window after stimulus onset. The spiking rates were normalized for each site (and session) by subtracting its mean response to the 25 normalizer images and then dividing by the standard deviation of its response over those normalizer images. This was done to compensate for day-to-day variations and allow meaningful comparisons across days and sites.

B Decoder training

Training and test images were randomly drawn from each domain (Figure 1) in 100 splits. Per split, we fit a linear ridge classifier trained on HVM images. Hyperparameter fitting was performed via leave-one-out cross-validation. Accuracies were averaged (mean) over all 100 splits.

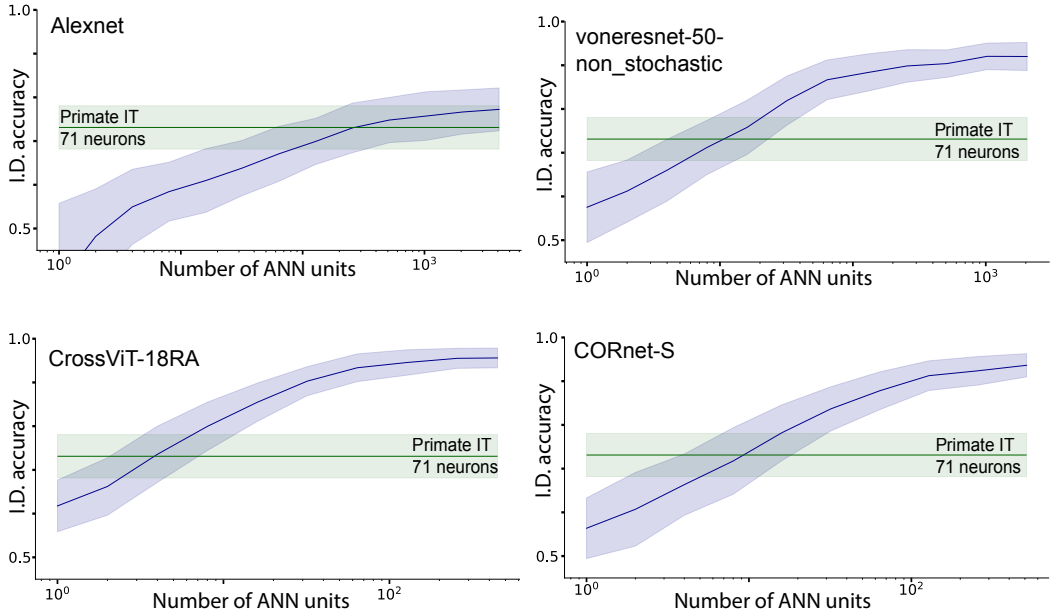

Figure B.1: **Scaling model units to primate neurons.** As in Figure 3A, for the remaining ANN models.

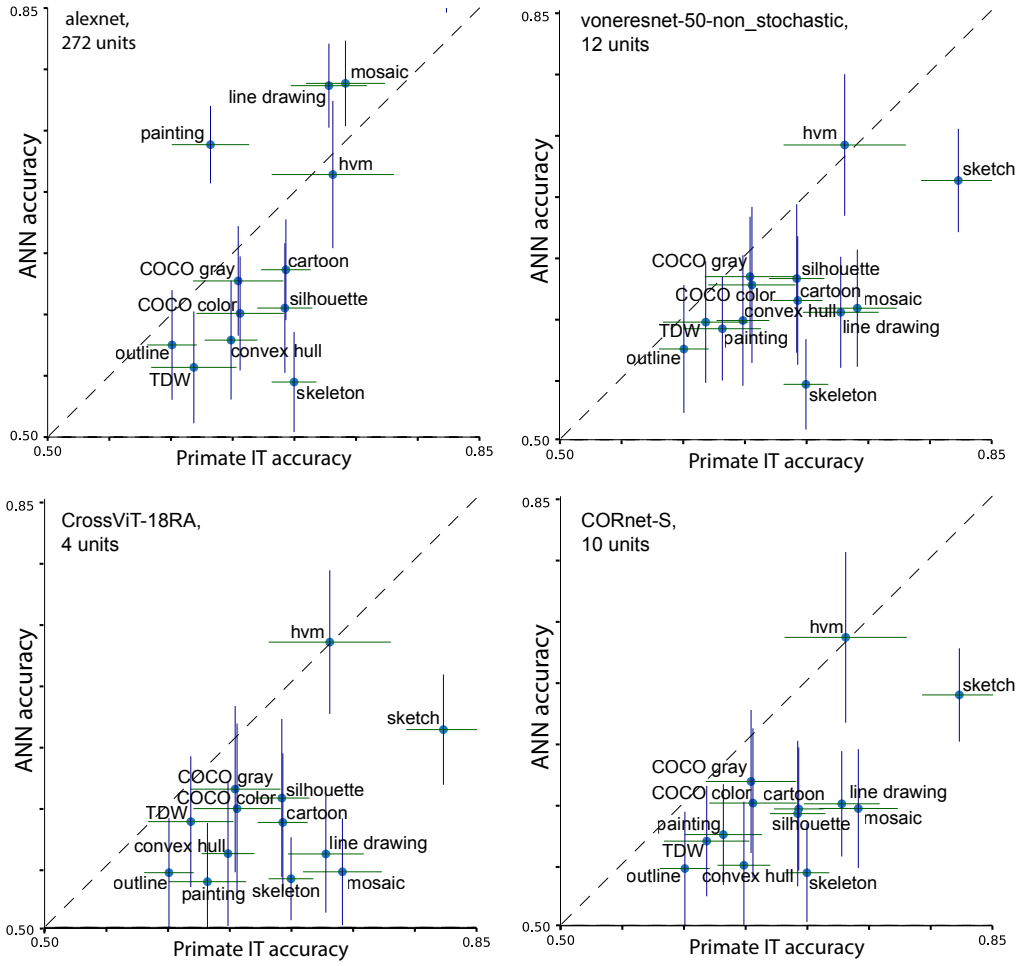

Figure B.2: **Alignment of ANN and primate IT generalization performance.** As in Figure 3B, for the remaining ANN models.