CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 093 November 2, 2018

Single units in a deep neural network functionally correspond with neurons in the brain: preliminary results

Luke Arend, Yena Han, Martin Schrimpf, Pouya Bashivan, Kohitij Kar, Tomaso Poggio, James J. DiCarlo and Xavier Boix
Center for Brains, Minds and Machines

Abstract

Deep neural networks have been shown to predict neural responses in higher visual cortex. The mapping from the model to a neuron in the brain occurs through a linear combination of many units in the model, leaving open the question of whether there also exists a correspondence at the level of individual neurons. Here we show that there exist many one-to-one mappings between single units in a deep neural network model and neurons in the brain. We show that this correspondence at the single-unit level is ubiquitous among state-of-the-art deep neural networks, and grows more pronounced for models with higher performance on a large-scale visual recognition task. Comparing matched populations—in the brain and in a model—we demonstrate a further correspondence at the level of the population code: stimulus category can be partially decoded from real neural responses using a classifier trained purely on a matched population of artificial units in a model. This provides a new point of investigation for phenomena which require fine-grained mappings between deep neural networks and the brain.

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

Single units in a deep neural network functionally correspond with neurons in the brain: preliminary results

Luke Arend, Yena Han, Martin Schrimpf, Pouya Bashivan, Kohitij Kar, Tomaso Poggio, James J. DiCarlo, and Xavier Boix

Center for Brains, Minds and Machines

Abstract

Deep neural networks have been shown to predict neural responses in higher visual cortex. The mapping from the model to a neuron in the brain occurs through a linear combination of many units in the model, leaving open the question of whether there also exists a correspondence at the level of individual neurons. Here we show that there exist many one-to-one mappings between single units in a deep neural network model and neurons in the brain. We show that this correspondence at the single-unit level is ubiquitous among state-of-the-art deep neural networks, and grows more pronounced for models with higher performance on a large-scale visual recognition task. Comparing matched populations—in the brain and in a model—we demonstrate a further correspondence at the level of the population code: stimulus category can be partially decoded from real neural responses using a classifier trained purely on a matched population of artificial units in a model. This provides a new point of investigation for phenomena which require fine-grained mappings between deep neural networks and the brain.

1 Introduction

Deep convolutional neural networks have recently seen considerable success as models of the primate ventral visual stream.[1–4] This is achieved via a modeling paradigm in which a convolutional neural network predicts single unit responses in the brain, when the brain and model are presented with identical stimuli. The approach is based on an assumption that a layer in the network model has an equivalent basis with a particular region in visual cortex. Specifically, the mapping from model to individual neurons in the brain occurs via a linear combination of units in the model: a set of stimuli are used to identify a linear combination of artificial units whose responses best predict those of a given real neuron in the brain; then the model is evaluated by its accuracy in predicting neural responses on a novel stimulus set. While this approach has shown a coarse correspondence between layers of deep network models and particular regions of visual cortex,[1,5] it remains unclear whether a more fine-grained correspondence could be established between individual units in the deep network model and neurons in the brain. Here we show that single units in a deep network model can predict neural responses in higher visual cortex. Furthermore, we show that this correspondence also holds at the level of the population code, in that a classifier trained on a set of matched artificial units can read out object category from a real neural population. Our method also allows us to identify which units in the model remain unvalidated by neural data.

2 Methods

2.1 Stimuli

The stimuli used were 2560 grayscale images of objects from eight categories, with eight exemplars per category. Exemplars were rendered on randomized natural backgrounds with high variation in location, size and pose. Example stimuli are shown in figure 1. We refer the reader to previous studies[3,1,6,5] using these stimuli for further details.

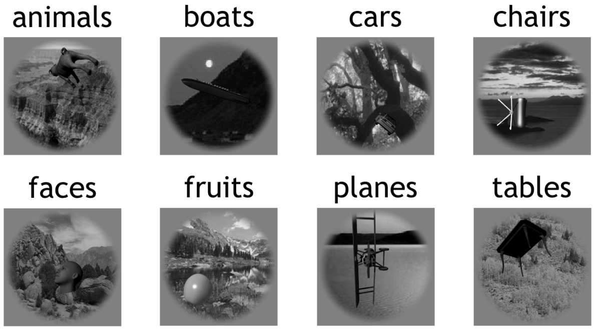


Figure 1: Example stimuli from each category.

Model	Top-1 accuracy (ImageNet)	Software framework
AlexNet[7]	0.577	PyTorch[8]
VGG-16[9]	0.715	keras[10]
NASNet-mobile[11]	0.74	TensorFlow-slim[12]
MobileNet-1.4[13]	0.75	TensorFlow-slim
Xception[14]	0.79	keras
DenseNet-121[15]	0.745	keras
DenseNet-169	0.759	keras
ResNet-50[16]	0.752	TensorFlow-slim
ResNet-101	0.764	TensorFlow-slim
ResNet-152	0.778	TensorFlow-slim
Inception-ResNet-v2[17]	0.804	TensorFlow-slim
Inception-v4[17]	0.802	TensorFlow-slim
NASNet-large[11]	0.827	TensorFlow-slim
PNASNet-large[18]	0.829	TensorFlow-slim

Table 1: Models used in the present study.

2.2 Neural recordings

The neural data were obtained via multi-electrode array recordings from areas V4 and IT of awake, alert macaque rhesus monkeys, yielding 88 sites from V4 and 168 sites from IT. Spike counts in the time window 70-170 ms after stimulus presentation were averaged across multiple presentations of each stimulus. We refer the reader to previous studies[3,1,6,5] using these data for further details.

2.3 Models

We used 14 state-of-the-art deep learning models implemented in three software frameworks. These are listed in table 1.

2.4 Metric

We propose a simple metric to compare individual units in a deep neural network with neurons in the brain. For each site in our set of neural recordings, we find the single unit (or, more generally, set of single units) whose responses across the stimulus set are most highly correlated. This procedure yields a population of "matched" units within a model which correspond, one-for-one, with the population recorded from the brain. Among this matched population, there are neuron-unit pairs which are correlated remarkably well (figure 2). We take the median correlation across the population as a single score for each model, which we denote here as "neural correlation."

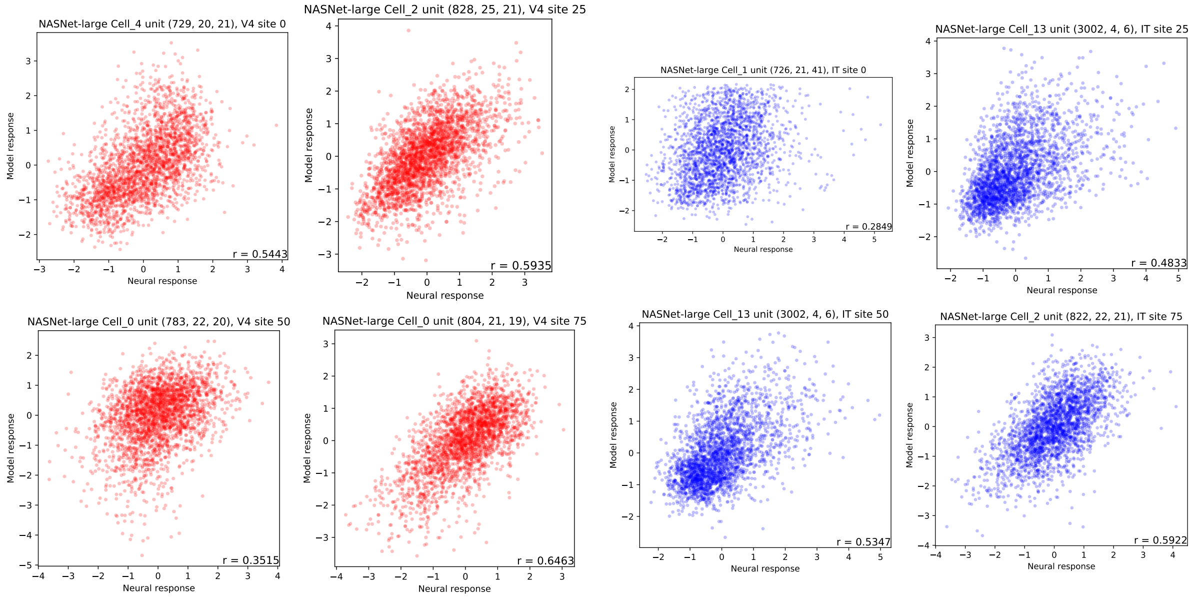


Figure 2: Examples of highly-correlated single units for V4 sites 0, 25, 50, 75 (left) and IT sites 0, 25, 50, 75 (right). Each point corresponds to one image in the stimulus set, with the x-axis showing the standardized (i.e. mean-subtracted and normalized to a standard deviation of 1) neural responses and the y-axis showing the standardized responses of a highly-correlated unit in the model. These units were identified within a high-performing network optimized for object recognition.

We consider the present approach to be promising but preliminary in light of a few considerations inviting additional investigation and validation. Note that units are matched with neurons via a simple correlation and selection of the most highly-correlated unit over a single stimulus set, rather than selecting matched pairs on one set of stimuli and re-evaluating correlations for these pairs on a novel set. While the procedure at hand is straightforward and simple, it may artificially favor larger or more expressive networks which can more readily overfit neural responses in the regime of limited stimuli. An experiment using various split sizes for selection and evaluation indicated that, for one network, our stimulus set was sufficiently large for neural correlation on withheld stimuli to plateau (figure S1); nonetheless, the conclusions presented here could potentially change if matched pairs are identified on one set of stimuli and then evaluated or analyzed using a novel set.

It should also be noted that the electrode array recordings used in the present work do not guarantee isolated activity from single neurons; previous work using this dataset has estimated that each electrode site captures potentials from one to three individual neural units.[1] Since a direct claim about one-to-one matching between model units and neurons would require neural recordings with isolated single units, throughout this paper we instead refer to a correspondence between single model units and single neural 'sites' (i.e. responses from a single electrode).

3 Results

We used the above metric to evaluate a set of 14 state-of-the-art deep neural networks (figure 3). The highest-scoring model overall was NASNet-large, with a neural correlation of 0.6224 for V4, and 0.4337 for IT. V4 correlation was remarkably high, showing that individual units in the model are well-tuned to match the response properties of neurons in V4. We also observed qualitatively that receptive fields computed for matched model units aligned well with independent physiological estimates of receptive field location for the corresponding V4 neural sites (figure S2). Models had greater difficulty capturing the complex responses patterns in IT, as correlations for IT sites, overall, were lower than those for V4. Neural correlation for regions IT and V4 increased together across models, with the NASNet-large model achieving the highest correlation for IT and the ResNet-152 model achieving the highest correlation for V4 (in the remainder of the text, results are reported for these two models). Consistent with previous reports,[1,19,3] neural correlation for both regions increases with performance on a recognition task. Neural correlation as a function of model size is shown in figure S3. Figure S4 shows a comparison between neural correlation for the single-unit matching metric presented here and a standard metric[1,3] that fits neural sites using regressed linear combinations of model units.

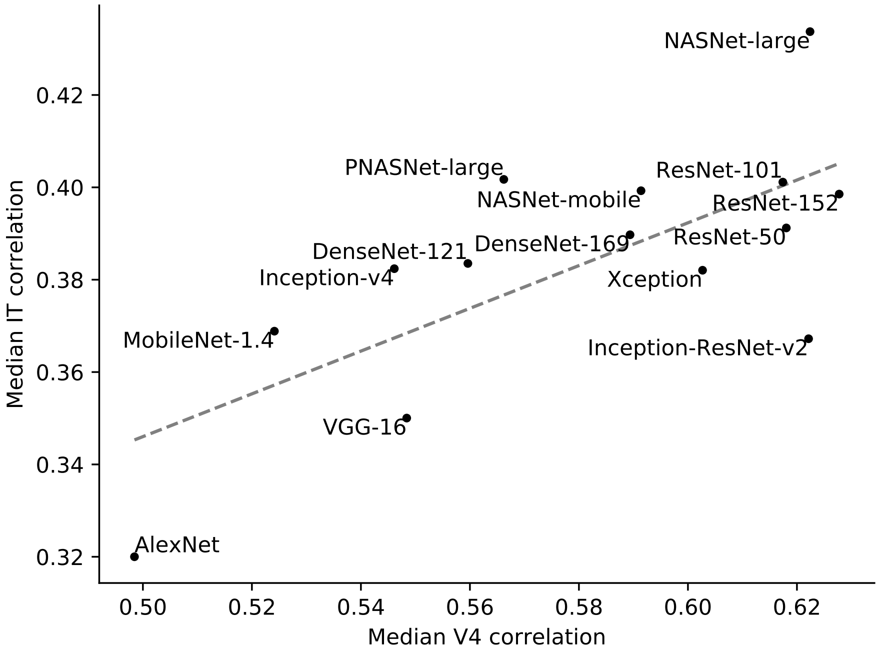


Figure 3: Results of the neural correlation metric for all 14 models. The y- and x-axes show the median (across sites) of the correlation of the best-matched units for each neural site in IT and V4, respectively.

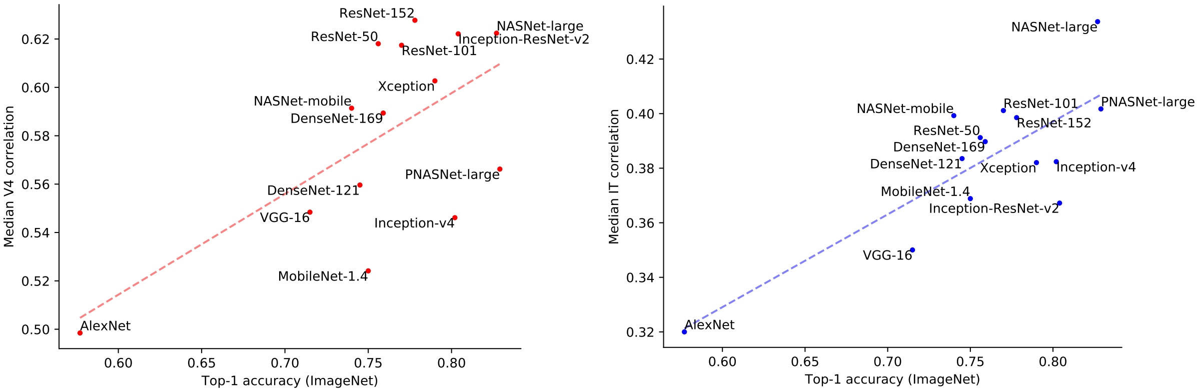


Figure 4: The relationship between neural correlation and task performance across models. The y-axis shows the median (across sites) of the correlation of the best-matched units for each neural site in V4 (left) and IT (right); the x-axis shows top-5 ImageNet classification accuracy.

Having matched up unit-for-neural site, we next investigated the extent to which this produced a further correspondence at the level of the population code. We trained a linear classifier to read out object category from the population responses, for each of the real and "matched" populations. Matched populations contained similar amounts of category information as the original neural population (figure 5). We found that stimulus category could be read out from neural responses in IT cortex using a classifier trained purely on the matched population of artificial units in the model (chance level = 0.125, decoder level = 0.3803). This was also true when decoding from the matched population in the model using a classifier trained on neural data. Taken together, these results suggest that the population-level response patterns for object category are at least partially shared between the model and the brain. Analysis of confusion matrices for the classifiers are shown in figures S5 and S6.

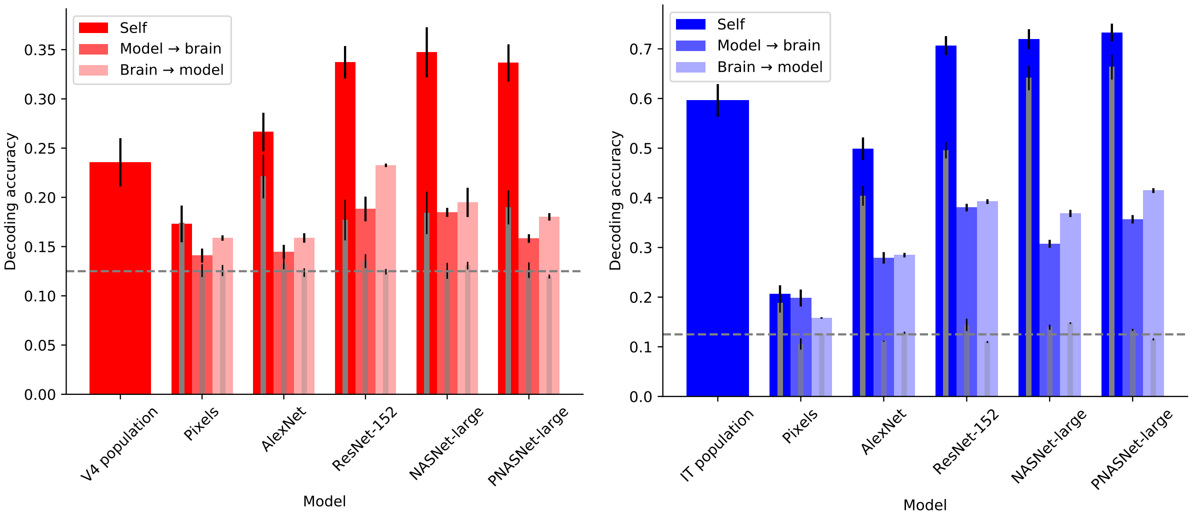


Figure 5: Decoding results for the neural population and various models. The y-axis shows test accuracy of a linear SVM classifier trained to decode object category from neural responses (for leftmost bar, "population") or from the responses of model units matched to the neural population. "Pixels" is a control model consisting of raw pixel activations in the input stimuli; pixels are selected according to the same procedure as is used to select model units. For each model, the three bars from left to right indicate the results of decoding under three conditions: (i) training and testing on the activations of matched model units, (ii) training on matched model unit activations and testing on neural population responses (iii) training on the neural population responses and testing on matched model unit activations. In condition (i), 90% of stimuli were used for training and 10% for testing. (The training and testing datasets for conditions (ii) and (iii) were the two different sets of responses—from brain and model—across all stimuli.) Error bars show standard deviation across multiple runs ($n = 25$ for condition (i); $n = 5$ for conditions (ii) and (iii)). Each run was cross-validated to select optimal training parameters using 10 random 90-10 splits of the training data. Dashed gray line indicates chance performance. Inset gray bars show, as a baseline, the same analysis performed instead with random (i.e. unmatched) model units sampled from the same layer as each original matched unit.

We next asked which layers in the model contained the matched units for neurons from each brain region. The results of this analysis are shown in figure 6 (left column). Generally, the matched population of units is distributed across the whole model. V4 neurons tend to be matched with units in the early layers of the model, while IT neurons are matched with units deeper in the model (figure S7). This suggests that some structural constraints are captured by the model (such as V4 roughly preceding IT in the feedforward mapping), but there is no overwhelming correspondence between any single layer in the network and a particular region of the brain. A priori, we expect this to be the case for very deep models because such models have many more stages of purely feedforward processing than are in the visual stream.

An open question when modeling the visual system using deep neural networks is the extent to which individual layers of the model align with particular areas of the ventral stream. We explored this by restricting the selection of matched units to a single layer and running the neural correlation procedure on every layer individually (figure 6, right column). We identified a "V4-like layer" and "IT-like layer" in each model as the single layer with the highest neural correlation for each of these two brain regions. In all models, the V4-like layer was found to precede the IT-like layer. Overall, restricting the metric to a single layer gives a noticeable (though not extreme) drop in neural correlation compared to evaluating neural correlation with access to all layers of the model (figure S8), supporting the notion that multiple layers of a model may be needed to capture the diversity of neuronal responses in each brain area.[4,20]

We sometimes found multiple layers which match to a brain region equally well, and these layers tended to occur in sequence. This is unsurprising if representational spaces within very deep networks are transformed only gradually from layer to layer, as has been noted previously.[21] A future study on the temporal dynamics of neural responses might investigate the hypothesis that many layers of a deep model serve as an "unrolled" approximation of a recurrent circuit.[22] An interesting case showing results for a ResNet architecture as depth (number of layers) increases is presented in figure S9.

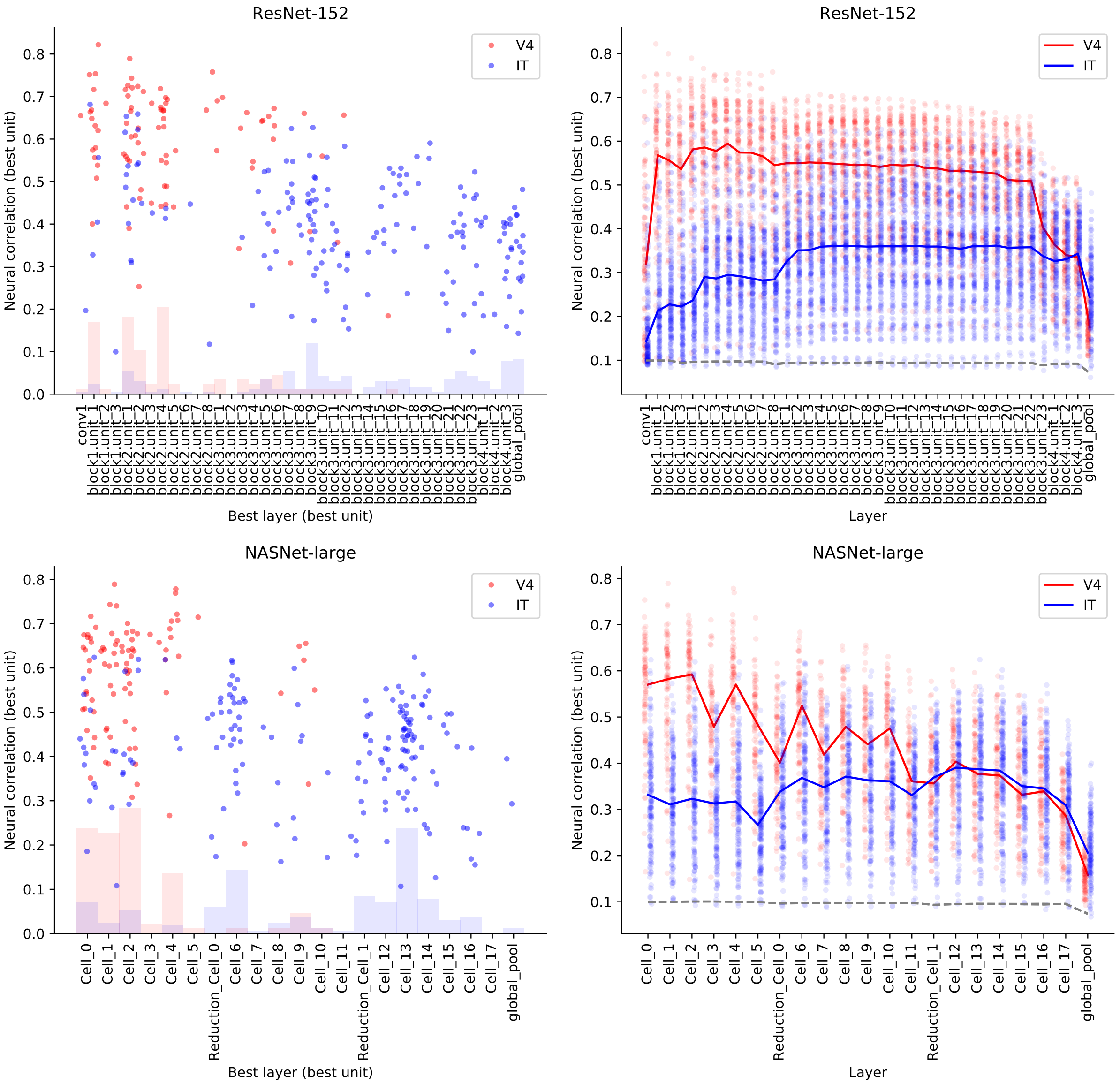


Figure 6: Results of layer analysis for the two models with highest neural correlation with V4 (top row, ResNet-152) and IT (bottom row, NASNet-large). The left column shows the layers of the model from which best-matched units were selected. Each point corresponds to the matched unit for a site from V4 (red) or IT (blue), with the y-value indicating the neural correlation of the unit and the x-axis indicating the layer of the model in which the unit was found. Transparent bars display as a histogram the distribution of matched units over model layers. The right column shows the result of running the matching procedure individually on each layer of the model. At each layer, the spread of points displays the correlations of the best-matched units in that layer for each V4 (red) or IT (blue) site; the solid line indicates the median of this distribution plotted across layers. The dashed gray lines (two, which are on top of each other), indicate the results of running the same analyses on a control model having equivalent layers which instead produces random, normally-distributed activations.

We went beyond simply considering the best-matched unit by investigating "runner-up" units—others in the model which also had high correlation with a given neural site. In particular, we looked at the distribution over correlation values of the top 100 units in the model matched to each site. The results of these analyses are shown in figure 7. Further analyses comparing this result for two models—one known to be redundant and another designed to be non-redundant—are shown in figure S10.

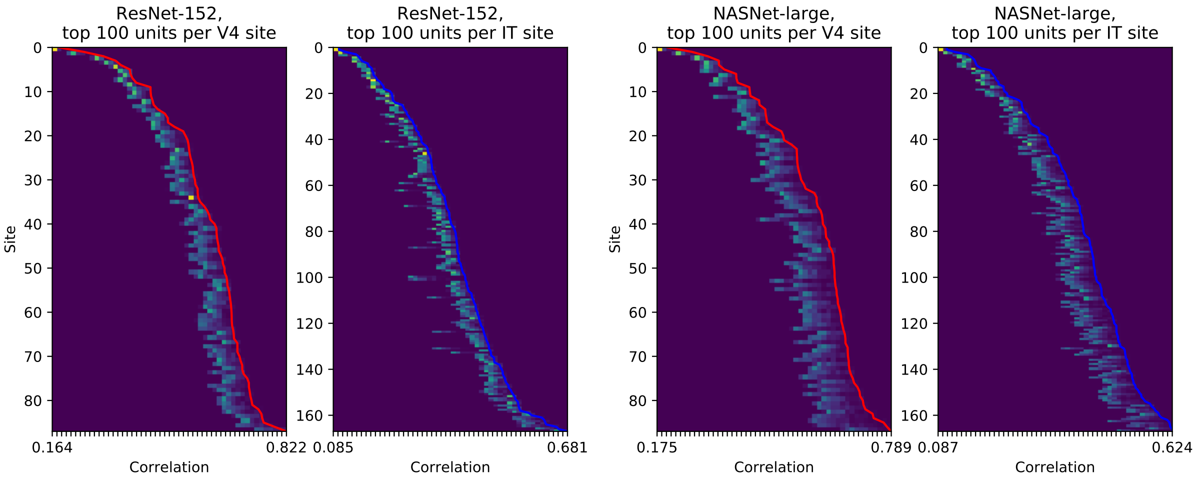


Figure 7: Correlation for the top-100 matching units for each site, shown for the two models having highest neural correlation with regions V4 (two leftmost subfigures, ResNet-152) and IT (two rightmost subfigures, NASNet-large). Each model has a subfigure for V4 (left) and IT (right). Each row of these images is read as a histogram plotted in intensity over correlation on the x-axis, showing the binned counts of correlation values for the top-100 matched units in the model for a single site. The solid line indicates the correlation of the best-matched (top-1) unit for each site; the rows are sorted by this value. Probability mass falling closer to the solid line indicates that more units, within the top 100, were near to the overall best unit found.

Next we asked—for V4- and IT-like layers—which particular feature maps (or 'kernels') and spatial locations within the convolutional layer produced the best-matching units for our neural population (figure 8). (We show the same analysis for the top 100 units in figure S11.) Frequently, a large number of sites from our neural recordings were matched to a single feature map. This is potentially explained by the fact that sites in the neural population are highly correlated amongst themselves (figure S12). Analyses of this sort are a potential advantage of the single-unit neural correlation metric presented here: they allow us to characterize precisely which aspects of the model are validated (or remain unvalidated) by the neural recordings we have (figure S13). Such a paradigm may prove useful for guiding future electrophysiological experiments, for instance, by searching for neurons in the brain which match units that remain unvalidated in the model.

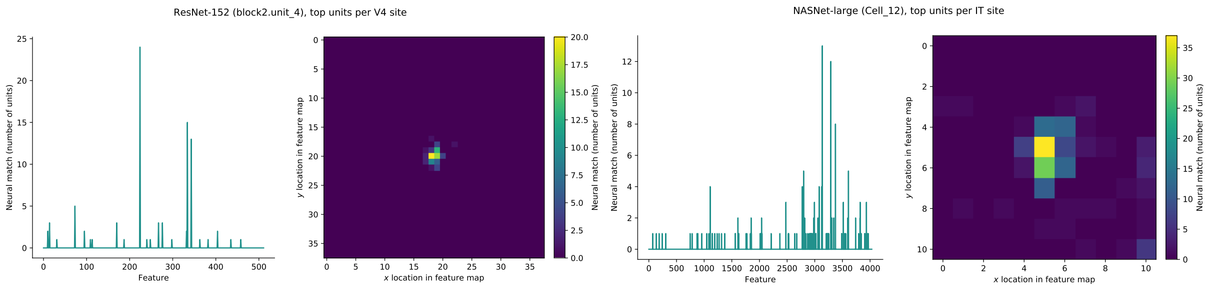


Figure 8: An investigation of which feature maps and spatial locations within a layer produce the best-matched units for the recorded population. The left subplot shows the result for the highest-scoring layer of the highest-scoring model in V4; the right subplot shows the same for area IT. The left side of each subplot shows a histogram of the number of neural sites matched to (any unit in) each feature map of the layer; the right side shows a histogram in intensity of the number of sites matched to (any feature in) each x or y spatial location of the layer.

We finally asked the question of whether neural correlation depended on the *particular* units in these models, or whether any layer with as many units (sometimes called "directions"[23]) in the same feature space would yield the same result. In other words, is there anything special about the natural basis formed by these particular units, or would any equivalent basis align equally well with neurons in the brain? To explore this, we transformed V4-like and IT-like layers with random rotations in feature space, and compared the neural correlations for these transformed layers (*i.e.*, artificially rotated bases) with the original, unrotated layers (the natural basis). The results of this analysis are shown in figure 9. In general, neural correlation was noticeably lower for rotated versions of a layer than for the unrotated layer.

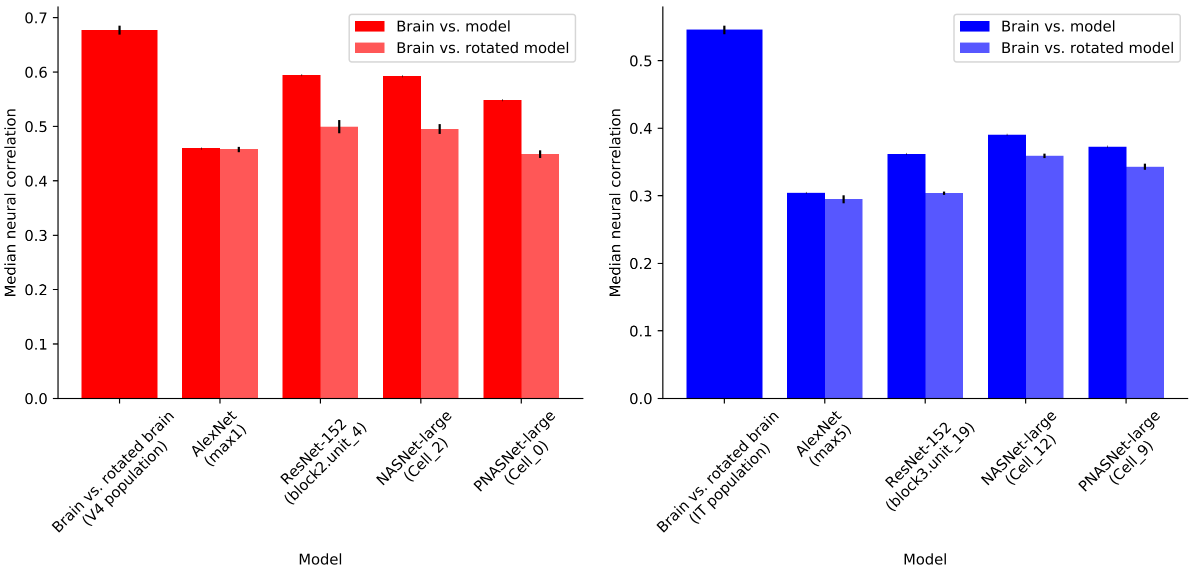


Figure 9: The effects of applying random rotations to the feature space of V4-like layers (left) or IT-like layers (right). For each model, the darker bar shows neural correlation for the highest-scoring layer, while the lighter bar shows neural correlation for the same layer transformed by a random rotation in feature space (error bars show standard deviation for $n = 10$ runs). For a layer with N feature kernels, a rotation matrix is randomly sampled from the group $SO(N)$ and applied to the N-dimensional vector of activations at each spatial location in the feature map, transforming the model's activations for each stimulus by a rotation in feature space. The matrix is sampled according to a procedure described previously.[24] For reference, the leftmost bar shows the result of applying a random rotation to the responses of the neural population (treating each neuron as a feature) and scoring neural correlation for the population against a rotation of itself.

Acknowledgements

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216).

Supplementary materials

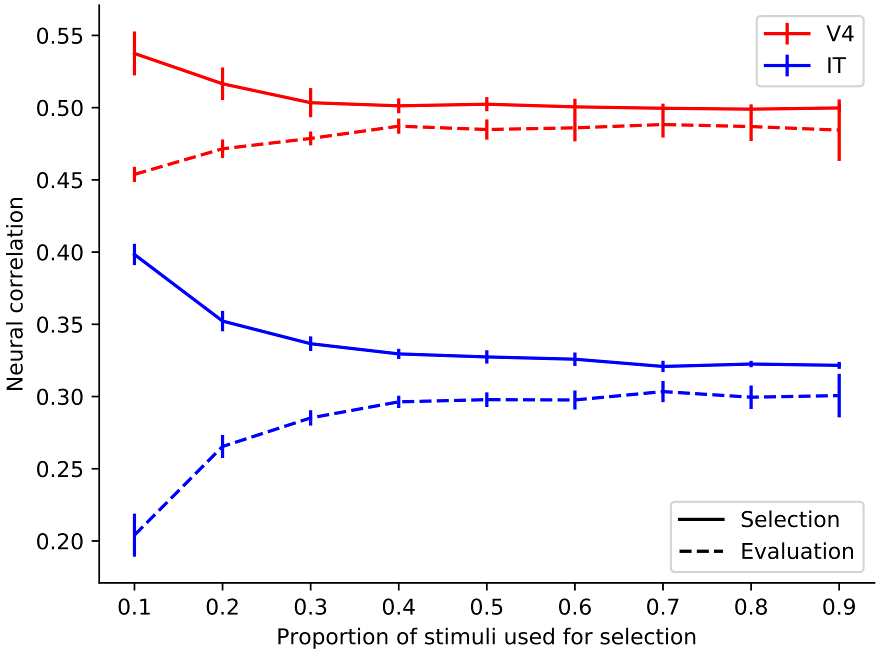


Figure S1: Neural correlation scores for V4 (red) and IT (blue) as the proportion of stimuli used for selection and evaluation are varied, for one network (AlexNet). Solid line indicates neural correlation score on the stimuli used to select matched pairs, and dashed indicates score for these selected pairs evaluated on the remaining (withheld) stimuli. Error bars show standard deviation across 25 runs with randomized splits.

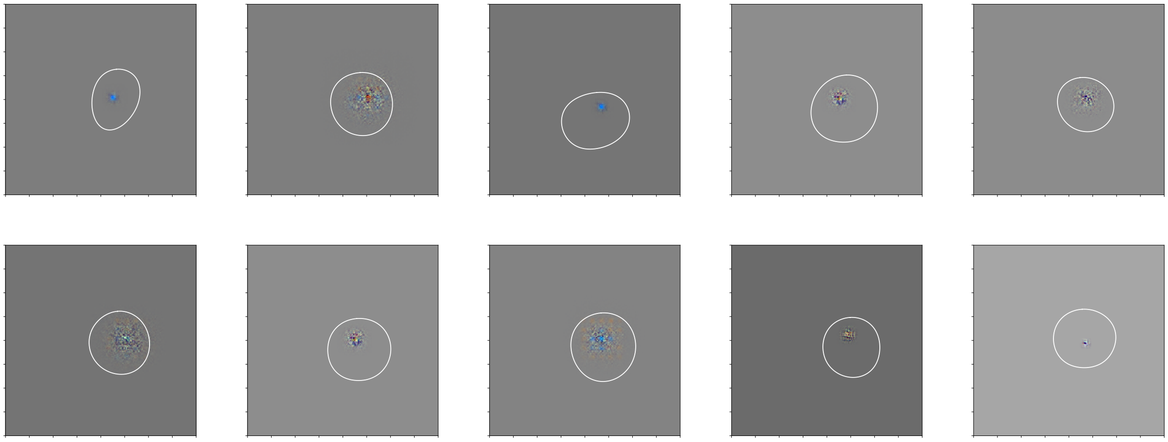


Figure S2: Receptive fields of example model units compared to physiologically estimated receptive fields for their corresponding sites in V4. Instantaneous receptive field patterns for model units were computed via a backward pass from the unit to the input image space to yield the visual gradient which maximized increase in activation (averaged across initializations from 64 random stimuli). Physiological receptive fields were estimated from neural responses to presentations of a white square of width 1° at each of 64 locations in an 8x8 grid covering 8° of the central visual field. Response maps were smoothed with a Gaussian kernel of bandwith 1° and normalized, then receptive fields were defined by the contours at which probability density equaled that of a normal distribution at 1 standard deviation. These physiological estimates of receptive field size can only be treated as an upper bound due to the spatial coarseness of the stimuli, so the model often suggests—or reveals—receptive field sizes to be considerably smaller. Locations align in all cases.

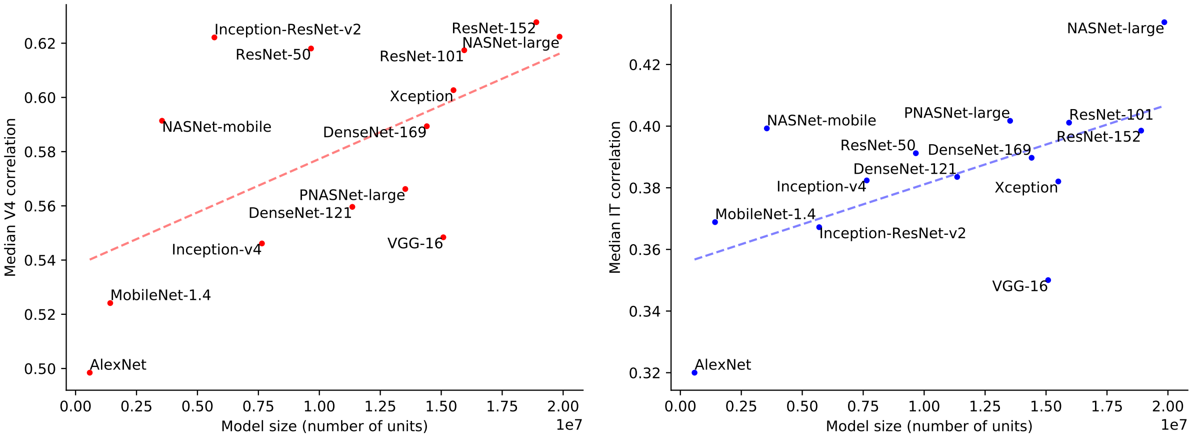


Figure S3: The relationship between neural correlation and model size, across models. The y-axis shows the median (across sites) of the correlation of the best-matched units for each neural site in V4 (left) and IT (right); the x-axis shows the number of model units.

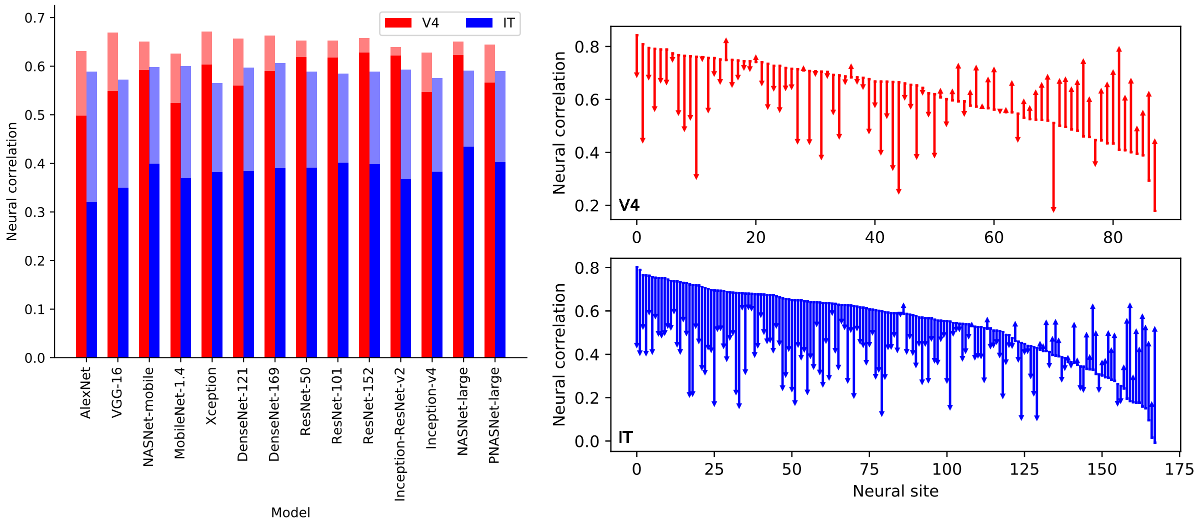
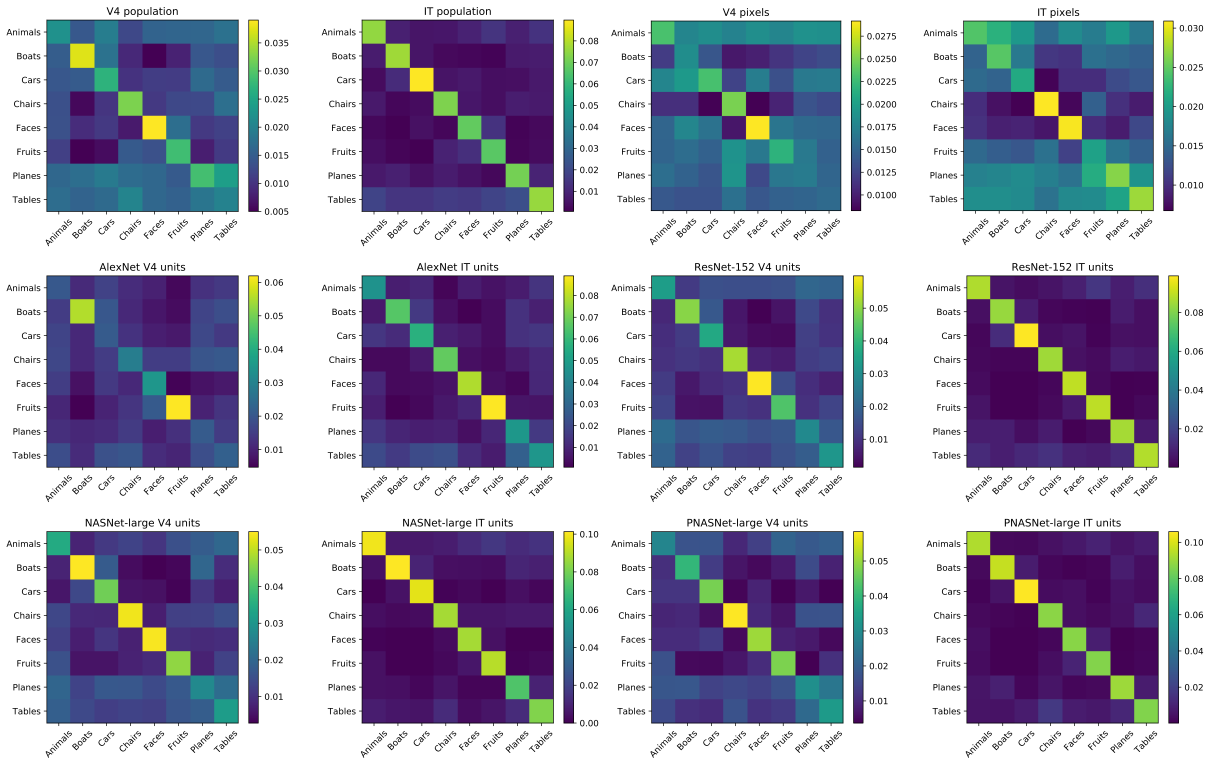


Figure S4: Comparison between the single-unit matching metric presented here and a standard metric that fits sites in each region using linear combinations of units from a model layer (results are from a recent report[3]). The left plot shows a comparison between the standard metric (faint bars) and single-unit metric (solid bars) across all models. Note that the single-unit metric can be seen as stringent special case of the standard metric (a linear combination of model units with a single nonzero weight); it thus yields lower scores overall. The right plot shows the change in neural correlation for each neural site when switching from the standard metric to the single-unit metric, for the two models with highest neural correlation with V4 (upper, ResNet-152) and IT (lower, NASNet-large). Sites are sorted by score on the standard metric. Scores for individual sites are sometimes higher with the single-unit metric; this is potentially explained by the fact that the standard metric is restricted to fitting all sites in each region using features from the same (single) model layer, while the single-unit metric is not—it may identify units matching sites in a region from any layer of the model. Note that this comparison is indirect in that the standard metric employs prediction on a holdout set, while the single-unit metric simply uses correlations (treated as a summary statistic) on a single dataset without holdout. Scores for the single-unit metric would be lower if it were used for prediction on a holdout set (see figure S1); nonetheless the comparison serves to give a sense for the difference in scores between the two procedures.


V4 population

IT population

V4 pixels

IT pixels

AlexNet V4 units

AlexNet IT units

ResNet-152 V4 units

ResNet-152 IT units

NASNet-large V4 units

NASNet-large IT units

PNASNet-large V4 units

PNASNet-large IT units

Figure S5: Confusion matrices, computed on the test set predictions, for the neural population (upper left) and each model used in the decoding analyses. Each adjacent pair of matrices shows the category-level confusion for V4 on the left, IT on the right. Confusion matrices are averaged across all 25 runs for the decoding analysis.

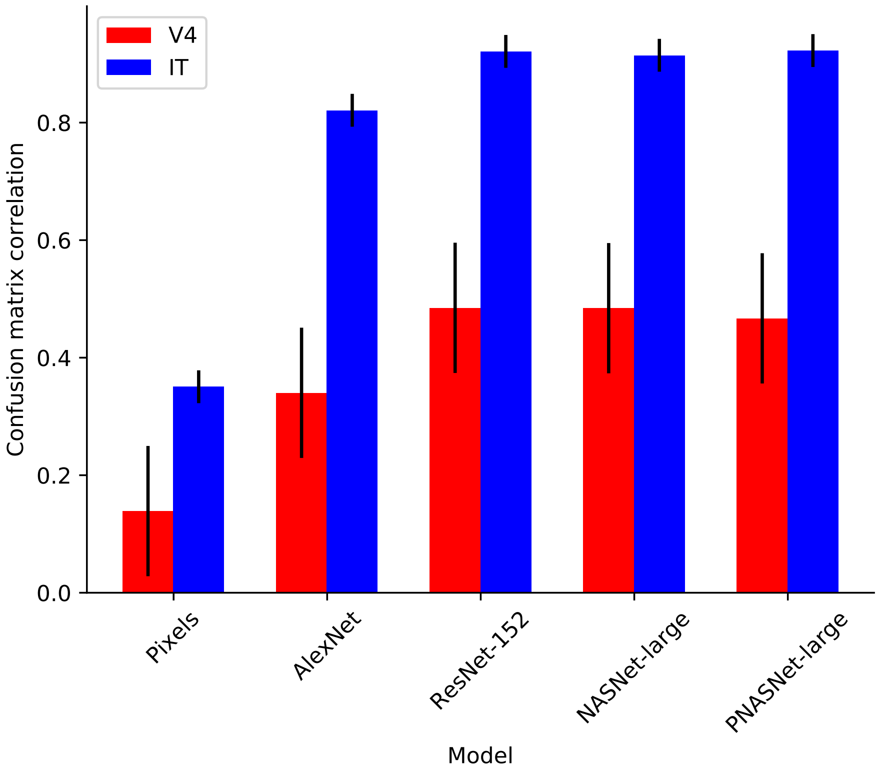


Figure S6: Correlation of confusion matrices for each model with the confusion matrices of the neural population, for V4 (red) and IT (blue). Error bars show standard deviation across all 25 runs.

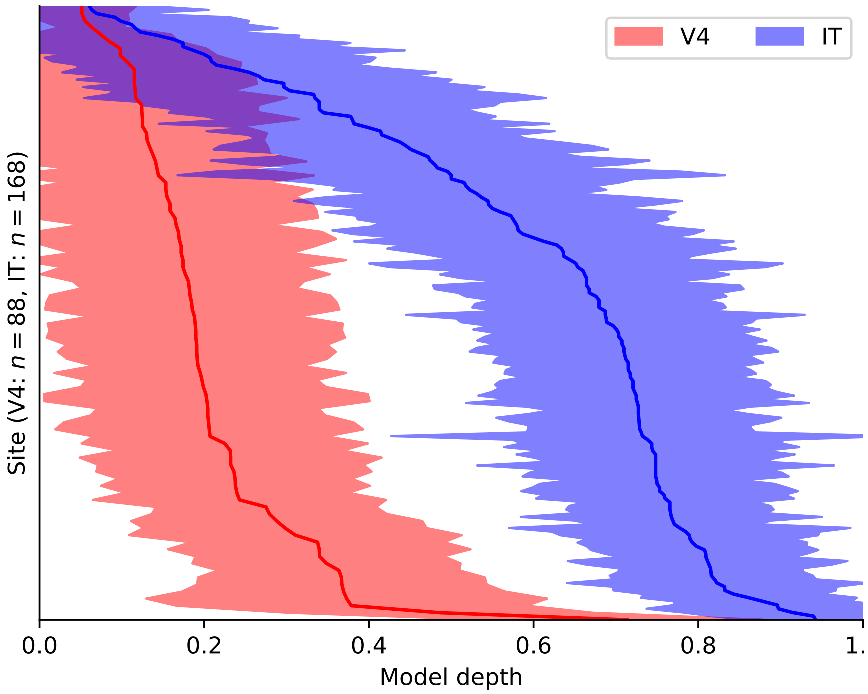


Figure S7: Here we consider the depth at which matching units for each neural site tend to be found, across all models used in the present study. As a simple and coarse measure of model depth, we enumerated the layers of each model with linearly spaced values between 0.0 (first layer) and 1.0 (last layer). For each site in V4 (red) and IT (blue), we then considered the median depth at which its best-matching unit was found in each model. This depth is plotted along the x-axis, with error bars indicating standard deviation across models. Neural site varies along the y-axis, with sites sorted by median depth.

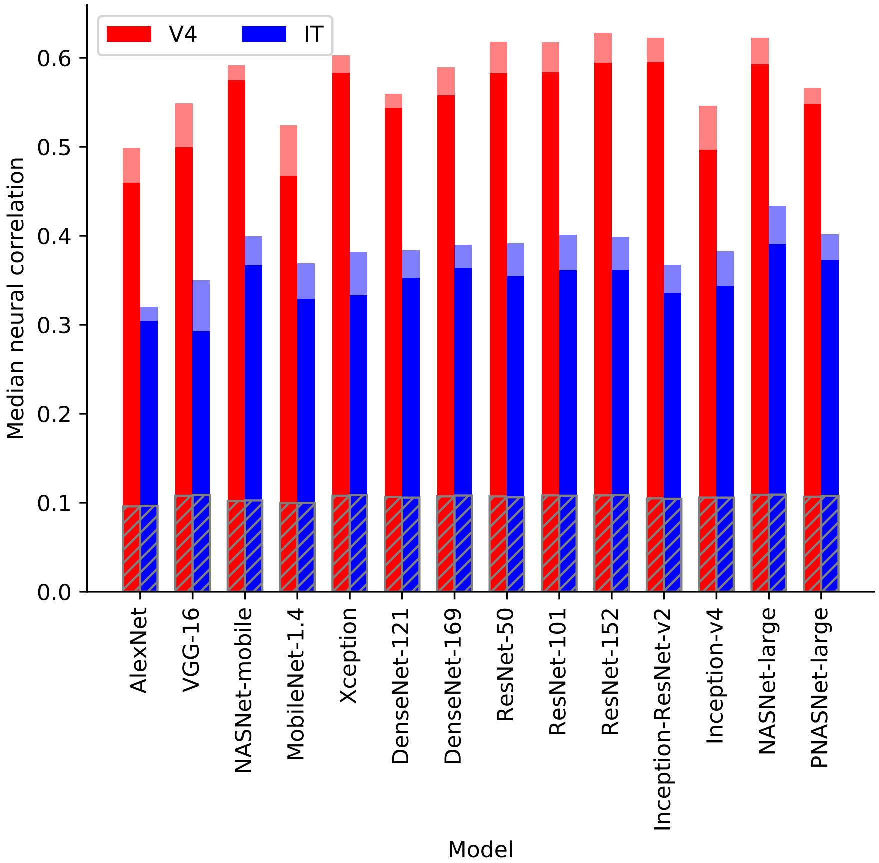


Figure S8: Comparison between our standard neural correlation metric (selecting matched units over the whole model; faint bars) and a single-layer restricted variant of the neural correlation metric (solid bars), for areas V4 (red) and IT (blue). The hatched gray bars indicate the results of running the same analyses on a control model having equivalent layers which instead produces random, normally-distributed activations.

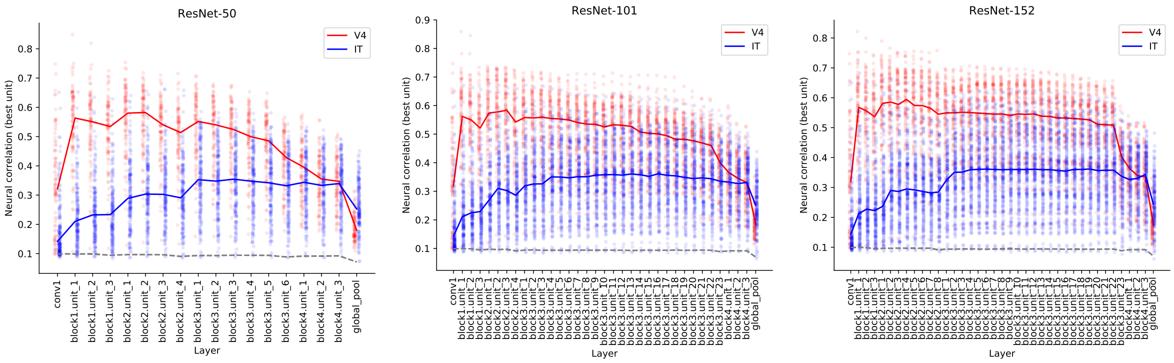


Figure S9: A side-by-side comparison of neural correlation across layers for the same architecture as depth increases. Layout is the same as in figure 6.

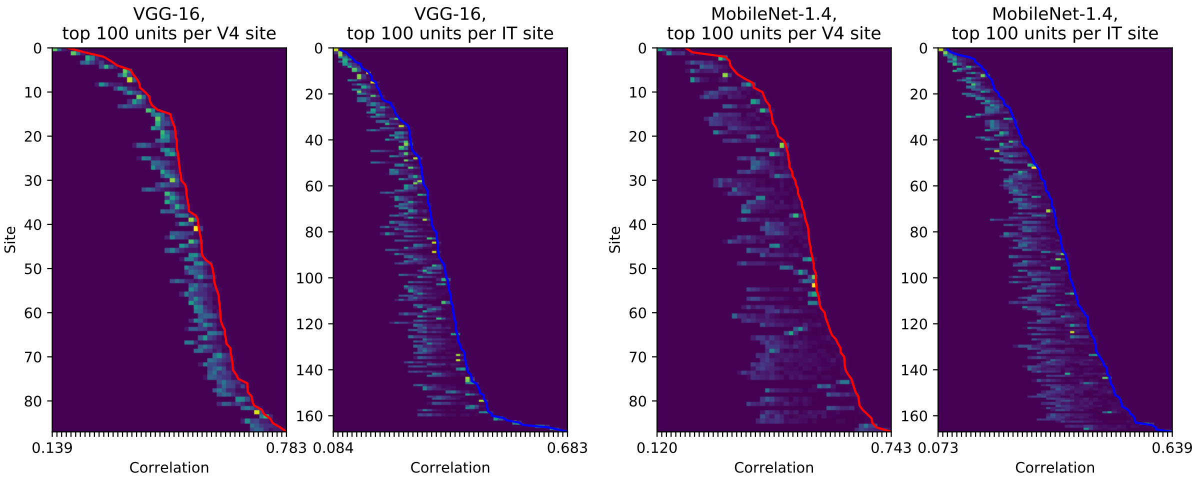


Figure S10: Correlations for top-100 matched units for each site, for a network which is known to be highly redundant (left, VGG-16) and a network which is optimized for efficient use of parameters (right, MobileNet-1.4). Layout is the same as figure 7.

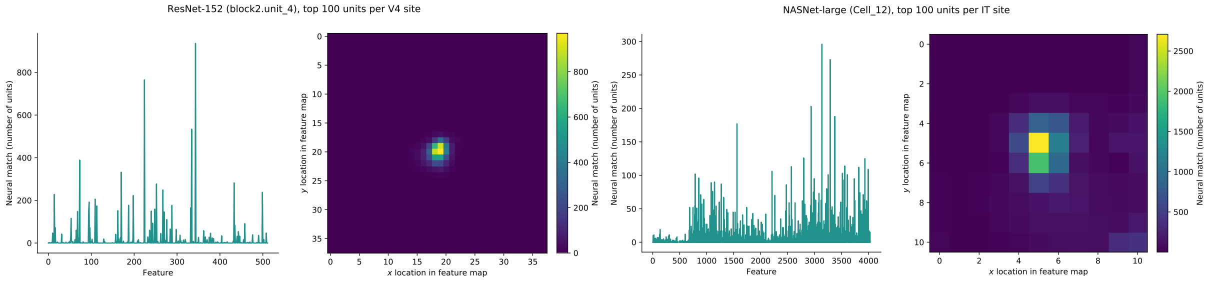


Figure S11: An investigation of which feature maps and spatial locations within a layer produce the top 100 matched units for each site in the recorded population. Other than including top 100 units rather than only the best unit for each site, layout is the same as in figure 8.

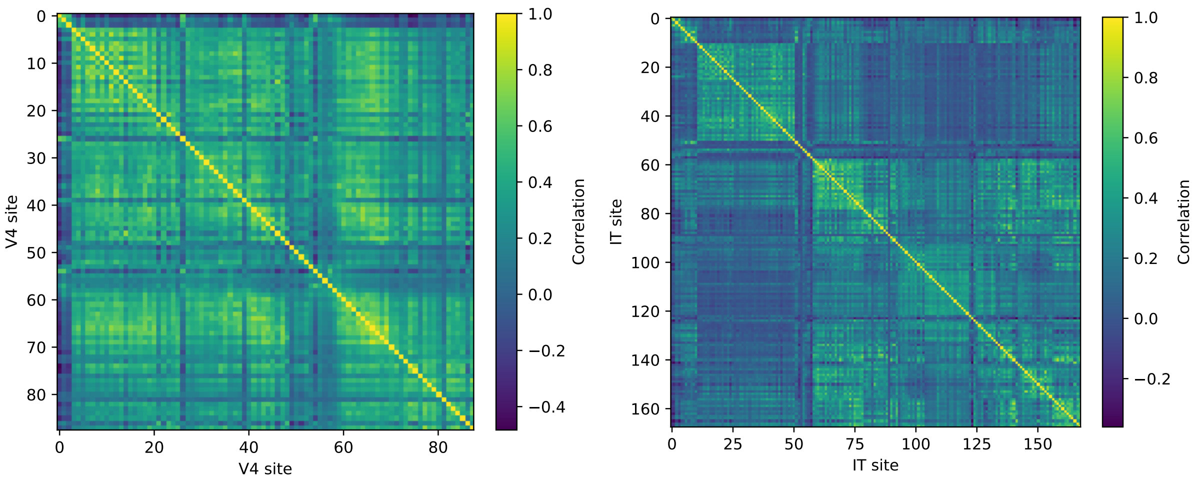


Figure S12: Correlation of neural responses across the stimulus set for all vs. all sites in the V4 (left) and IT (right) populations.

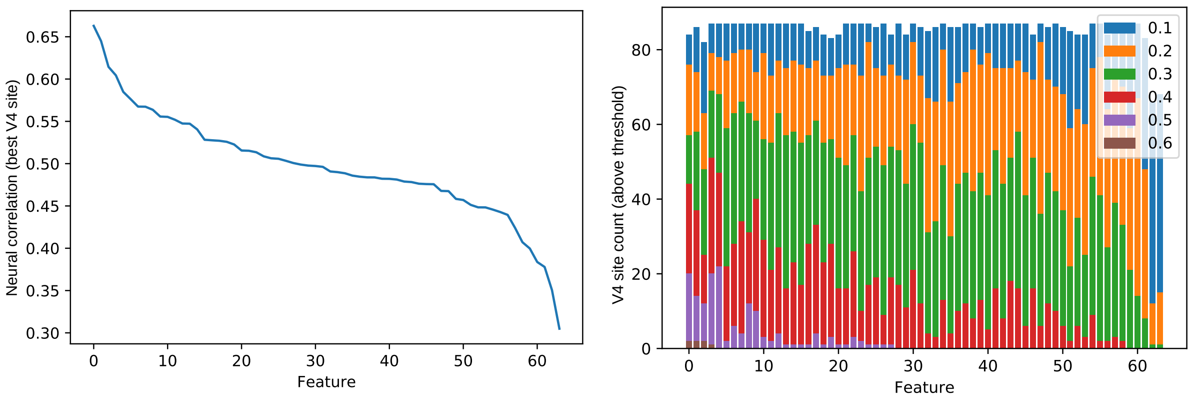


Figure S13: A prototypical analysis of which feature maps in the model are validated or unvalidated by neural data. Conceptually, instead of searching for matches to real neurons among model features, we search for matches to model features among our recorded neurons. The left figure shows the neural correlation of the best-matched V4 site (y-axis) for each feature map in a given layer of a model (x-axis, sorted by correlation). When matching each neural site to a feature map, we match to the most highly-correlated unit in the given feature map for that site. The right figure shows, for each feature map along the x-axis, the number of V4 sites (out of 88) whose correlation with that feature map is above a given threshold, indicated in color. This analysis is performed on the first max-pooling layer of the AlexNet model. Feature indices are the same in both plots.

References

[1] Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci* **111**, 8619–24 (2014).

[2] Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci* **19**, 356–65 (2016).

[3] Schrimpf, M. *et al.* Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv* (2018).

[4] Kubilius, J. *et al.* CORnet: Modeling the neural mechanisms of core object recognition. *bioRxiv* (2018).

[5] Hong, H., Yamins, D. L. K., Majaj, N. J. & DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat Neurosci* **19**, 613–22 (2016).

[6] Majaj, N. J., Hong, H., Solomon, E. A. & DiCarlo, J. J. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *J Neurosci* **35**, 13402–18 (2015).

[7] Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, 1097–1105 (2012).

[8] Paszke, A. *et al.* Automatic differentiation in PyTorch (2017).

[9] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR* **abs/1409.1556** (2014).

[10] Chollet, F. *et al.* Keras. `https://keras.io` (2015).

[11] Zoph, B., Vasudevan, V., Shlens, J. & Le, Q. V. Learning transferable architectures for scalable image recognition. *CoRR* **abs/1707.07012** (2017).

[12] Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems. `http://tensorflow.org/` (2015).

[13] Howard, A. G. *et al.* MobileNets: Efficient convolutional neural networks for mobile vision applications. *CoRR* **abs/1704.04861** (2017).

[14] Chollet, F. Xception: Deep learning with depthwise separable convolutions. *CoRR* **abs/1610.02357** (2016).

[15] Huang, G., Liu, Z. & Weinberger, K. Q. Densely connected convolutional networks. *CoRR* **abs/1608.06993** (2016).

[16] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *CoRR* **abs/1512.03385** (2015).

[17] Szegedy, C., Ioffe, S. & Vanhoucke, V. Inception-v4, Inception-ResNet and the impact of residual connections on learning. *CoRR* **abs/1602.07261** (2016).

[18] Liu, C. *et al.* Progressive neural architecture search. *CoRR* **abs/1712.00559** (2017).

[19] Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V. & McDermott, J. H. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**, 630–644.e16 (2018).

[20] Kalfas, I., Vinken, K. & Vogels, R. Representations of regular and irregular shapes by deep convolutional neural networks, monkey inferotemporal neurons and human judgments. *PLOS Computational Biology* **14**, 1–26 (2018).

[21] Cadena, S. A. *et al.* Deep convolutional models improve predictions of macaque V1 responses to natural images. *bioRxiv* (2017).

[22] Liao, Q. & Poggio, T. A. Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *CoRR* **abs/1604.03640** (2016). `1604.03640`.

[23] Morcos, A. S., Barrett, D. G. T., Rabinowitz, N. C. & Botvinick, M. On the importance of single directions for generalization. *ArXiv e-prints* (2018). `1803.06959`.

[24] Zhou, B., Bau, D., Oliva, A. & Torralba, A. Interpreting deep visual representations via network dissection. *CoRR* **abs/1711.05611** (2017).