Models of the ventral stream that categorize and visualize images

Elijah Christensen[1], Joel Zylberberg[2]

[1]Department of Physiology and Biophysics, University of Colorado Anschutz Medical Campus
[2]Learning in Machines and Brains Program, CIFAR, Toronto, ON Canada

Abstract

A widely held view in visual systems neuroscience is that the ventral stream (VS) of mammalian visual cortex maps visual inputs onto firing patterns that specify object identities. Recent experiments, however, showed that information about object positions, sizes, etc. is encoded with increasing explicitness along this pathway. Here, we show that computational models that identify objects while *also* retaining enough information about the input image to allow its reconstruction, can provide a better description of how primate VS represents this *category-orthogonal* information, than do models that "just" identify objects. A thorough understanding of VS computations might thus require considerations beyond object recognition.

Main

The ventral stream (VS) of visual cortex begins in primary visual cortex (V1), ends in inferior temporal cortex (IT), and is essential for object recognition. Accordingly, the long-standing belief in the field is that the ventral stream could be understood as mapping visual scenes onto neuronal firing patterns that represent object identity[1]. Supporting that assertion, deep convolutional neural networks (DCNN's) trained to categorize objects in natural images develop intermediate representations that resemble those in primate VS[2-5]. However, several experimental findings appear at odds with the object recognition hypothesis. VS and other visual areas are also engaged during visualization of both prior experience and novel scenes[6,7], suggesting that the VS can *generate* visual scenes, in addition to processing them as inputs. Furthermore, non-categorical information, about object positions[8], sizes, etc. is also represented with increasing explicitness in late VS areas V4 and IT[9]. This is not necessarily expected in a "pure" object recognition system, as the non-categorical information is not necessary for the categorization task. Thus these recent findings challenge notion that ventral stream is purely an object recognition system, and raise the question: What computational objective best explains VS physiology[10]?

To address this question, we pursued a recently-popularized approach and trained deep neural networks to perform different tasks: we then compared the trained neural networks' responses to image stimuli to those observed in neurophysiology experiments[3-5,9], to see which tasks yielded models that best matched the neural data. We trained our networks to perform one of two visual tasks: a) recognize objects; or b) recognize objects while *also* retaining enough information about the input image to allow its reconstruction. We studied the evolution of categorical and non-categorical information representations along the visual pathway within these models, and compared that evolution with data from monkey VS. Our main finding is that neural networks optimized for task (b) provide a better match to the representation of non-categorical information in the monkey physiology data than do those optimized for task (a). This suggests that a full understanding of visual ventral stream computations might require considerations other than object recognition.

Computational Models

To identify the degree to which different computational objectives describe ventral stream physiology, we optimized computational neural network models for different objectives, and compared them to neural recordings from the primate ventral stream. Each computational model was constructed out of a series of layers of artificial neurons, connected sequentially. The first layer takes as input an image $\boldsymbol{x}$ and at the final layer outputs a set of neuronal activities that represent the visual scene input (Fig 1B), including object identity. We refer to this output as the *latent representation*. The input images, $\boldsymbol{x}$, consisted of images of clothing articles superimposed over natural image backgrounds (see Methods). Each image used a single clothing article rendered in a randomly chosen position and placed over a natural image background (Fig. 1A).

The models each had a total of three layers of processing (corresponding to cortical areas V1, V2, and V4) between their inputs and these latent representations; the latent representations correspond to area IT, for reasons we discuss below. The visual inputs to the model had normalized luminance values, mimicking the normalization observed in thalamic inputs to V1[11]. The connectivity between neurons in each layer (and the artificial neurons' biases) were optimized within each model, so as to achieve the specified objective (see Methods). We repeated this process for two different objectives, yielding two different types of models.

The first type of model was optimized strictly for object recognition: the optimization maximized the ability of a linear decoder to determine the identity of the clothing object in the visual scene from the latent representation. (This mirrors the observation that neural activities in area IT can be linearly decoded to recover object identity[12]). The second type of model was optimized for two tasks in parallel: the ability of a linear decoder to determine object identify from latent representation, *and* the ability of a decoder to reconstruct the object from the latent representation. (See Methods for details about the optimization procedure). We repeated this procedure with both convolutional, and non-convolutional neural network architectures, yielding a total of four models (Fig 1C).

In all cases, the models were optimized using sets of images containing randomly sampled objects, until their object classification performance saturated on a set of held-out validation images. Good performance on the categorization task was obtained in all models (Fig 1D). Having developed models optimized for these different objectives, we could evaluate how well each model matched observations from primate VS, and use that comparison to determine which computational objective provides the best description of primate VS.

Electrophysiology Comparisons

To compare our neural network models to ventral stream physiology, we used the experimental data from a previously-published study[9,12] (see methods and ref[9,12] for details). These data consisted of electrode array recordings from areas V4 and IT of monkeys that were viewing images; many neurons in each area were simultaneously observed. Within these data, we assessed each neuron's selectivity for object identify, and for category-orthogonal image properties (e.g. horizontal object position), as in Hong et al[9] (see methods). We performed this analysis for the monkey data, and for the artificial neurons in each layer of each of our computational models. We then compared the trends in image property selectivity displayed by non-human primate VS neurons and units from each of our models along the visual processing pathway.

In the primate VS, selectivity for both categorical and category-orthogonal scene attributes increased along the ventral stream (Fig 2A), as reported by Hong et al[9]. This indicates that both types of attributes are more explicitly represented in progressively deeper ventral stream areas.

Within our computational models, those models optimizing the composite objective showed the same trends observed in primate ventral stream neurons (Fig 2C, 2E): both category and category-orthogonal properties of the visual scene are represented more explicitly with each subsequent layers of the model. This observation persisted for both the convolutional and the non-convolutional architectures. For contrast, models optimized solely for object recognition (without the image reconstruction component of the objective function) did not show consistent increases in position selectivity along the visual pathway (Fig 2B, 2D). Again, this observation held for both convolutional and non-convolutional model architectures.

Thus, models optimizing the composite objective function qualitatively recapitulate the trends in neuronal selectivity along the visual pathways better than do models optimized strictly for object recognition. This observation motivated us to quantify how well each model matched the primate VS data. To achieve this goal, we performed the following analysis on each computational model. First, we used unpaired t-tests to estimate the probability that there is no difference in object category selectivity between the primate IT data and the model's latent representation. We then performed a t-test comparing the primate V4 category selectivity to the corresponding layer of the computational model. Next, we performed t-tests comparing the horizontal, and vertical, position selectivities in primate V4 and IT to the corresponding layers of the computational model. This procedure yielded 6 p-values, describing the probability that the model matched each of these attributes observed in the primate VS. Finally, we used Fisher's method[13] to combine those 6 p-values into a single number, that quantified the likelihood of there being no difference between the computational model and the primate VS.

Comparing these likelihood values, we found that the convolutional models overall provided better descriptions of the primate VS than did the non-convolutional ones (i.e., they had higher likelihood values), and that the best model overall was the convolutional neural network optimized for the composite classify-and-reconstruct objective (See Supplemental Fig. 1).

Noise Robustness

We found that the convolutional model, optimizing the composite objective (classify-and-reconstruct) best matched the depth-dependent increase in position selectivity seen in single unit activities recorded from primate ventral stream. This led us to ask whether there might be functional benefits for networks optimizing this composite objective function, as compared with ones that are just trained to classify their inputs.

Further motivating this question, we note that previous work has shown that convolutional neural networks optimized for object recognition tend to perform poorly on object recognition tasks when the images are corrupted by noise. Specifically, classification performance has been seen to decrease significantly when networks are evaluated under noise conditions even marginally different from the conditions under which it was trained[14]. This is different from the primate visual system, where object recognition performance is more robust to image noise, leading us to speculate that the convolutional networks trained for the composite classify-and-reconstruct task – which provide the best match to primate VS data – might have classification performance that is more robust to image corruption than do the networks trained purely for object recognition.

To test that hypothesis, we took each of our previously trained models, and measured their accuracy at categorizing the clothing objects in test images corrupted by increasing levels of additive pixel noise (see methods). Similar to previous work, the convolutional model trained purely for object recognition showed a decrease in performance as the noise level increased. For the convolutional model trained on the composite task, the decrease in performance with increasing noise level was less severe. This suggests that, consistent with our hypothesis, there is a functional benefit to systems optimizing the composite objective over "pure" object recognition systems: their object recognition performance is more robust to noise.

The same finding also holds for the non-convolutional model architectures, and they are overall more robust to image noise than are the convolutional ones. We repeated this analysis with multiplicative (instead of additive) pixel noise (see Supplemental Fig. 2) and demonstrate that our findings can be generalized across multiple noise types.

Discussion

Here we report evidence that convolutional neural networks (DCNNs) optimizing a two-part composite objective (recognize and visualize) describe the depth-dependent evolution of categorical and non-categorical information in primate VS better than do networks optimized for object recognition alone. This is unexpected, as prior work posits that networks optimized strictly for object recognition should form the best models of primate VS.[2,4,9,10] Our results suggest that the evolution of category-orthogonal information along the visual pathway could require a different functional explanation. Moreover, consistent with previous work,[2,4,9,10] our CNNs optimized for image classification resemble primate VS more closely than do non-convolutional models optimizing the same objective.

Our findings may help reconcile discrepancies between the object recognition hypothesis of VS and results which appear at odds with this interpretation,[6,7,15,16] for example the finding that primate VS explicitly retains information not useful for object recognition experiments tested previously.[9] The composite objective promotes retention of both category and category-orthogonal information because both are necessary to reconstruct the stimulus.

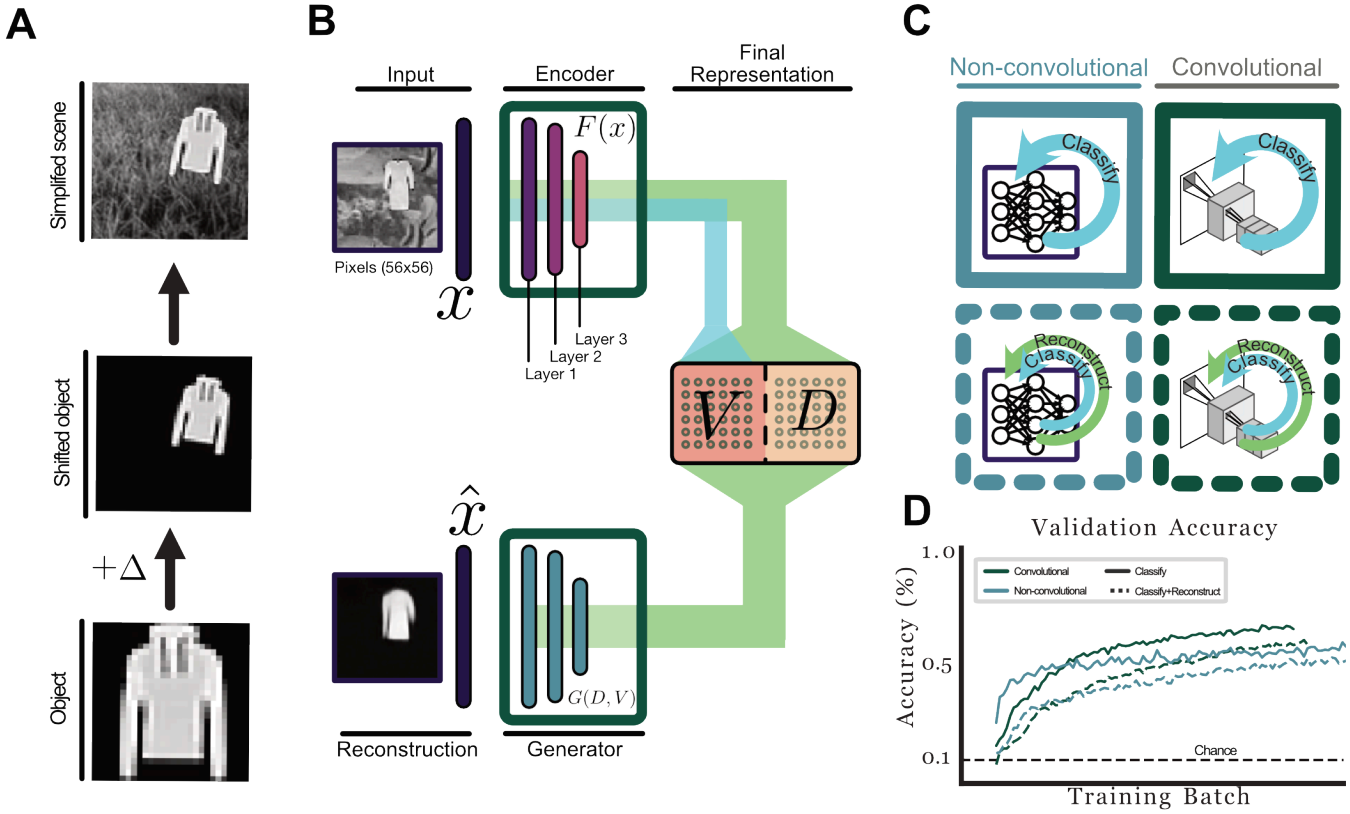
Importantly, we used a different method to compare our neural networks to the primate VS than have previous studies that compared the representational dissimilarity matrices (RDMs) for their models, with those of the primate VS.[2,4,9,10] While RDMs assay the (dis)similarity in how different images are represented by the models, or primate VS, recent work suggests RDM analysis may be insufficient as a universal metric of model similarity[17]; especially when the model cannot be trained using identical image datasets (as in our case). Instead, our approach was to focus on the depth-dependent evolution of neuronal selectivity to categorical and non-categorial variations in the input images. Our conclusion -- that an objective other than pure object categorization could best describe the computations in primate VS – differ from prior studies and further suggest that aspects of visual computation are not fully captured by RDM analysis.

Furthermore, our findings suggest noise tolerance as another independent explanation for why the VS might use a composite computational objective. VS classification accuracy measured in humans tolerates noise corrupted images much better than DCNNs optimized for image classification alone[14]. In contrast, convolutional model's optimizing the composite objective demonstrate better noise tolerance compared to identical models trained solely for classification (Fig 3). Importantly, improved noise tolerance occurs without having to augment training images with noise. These findings complement the expanding body of work to explain the neuronal computations in visual processing and have applications in the computer vision models that emulate them.

Acknowledgements

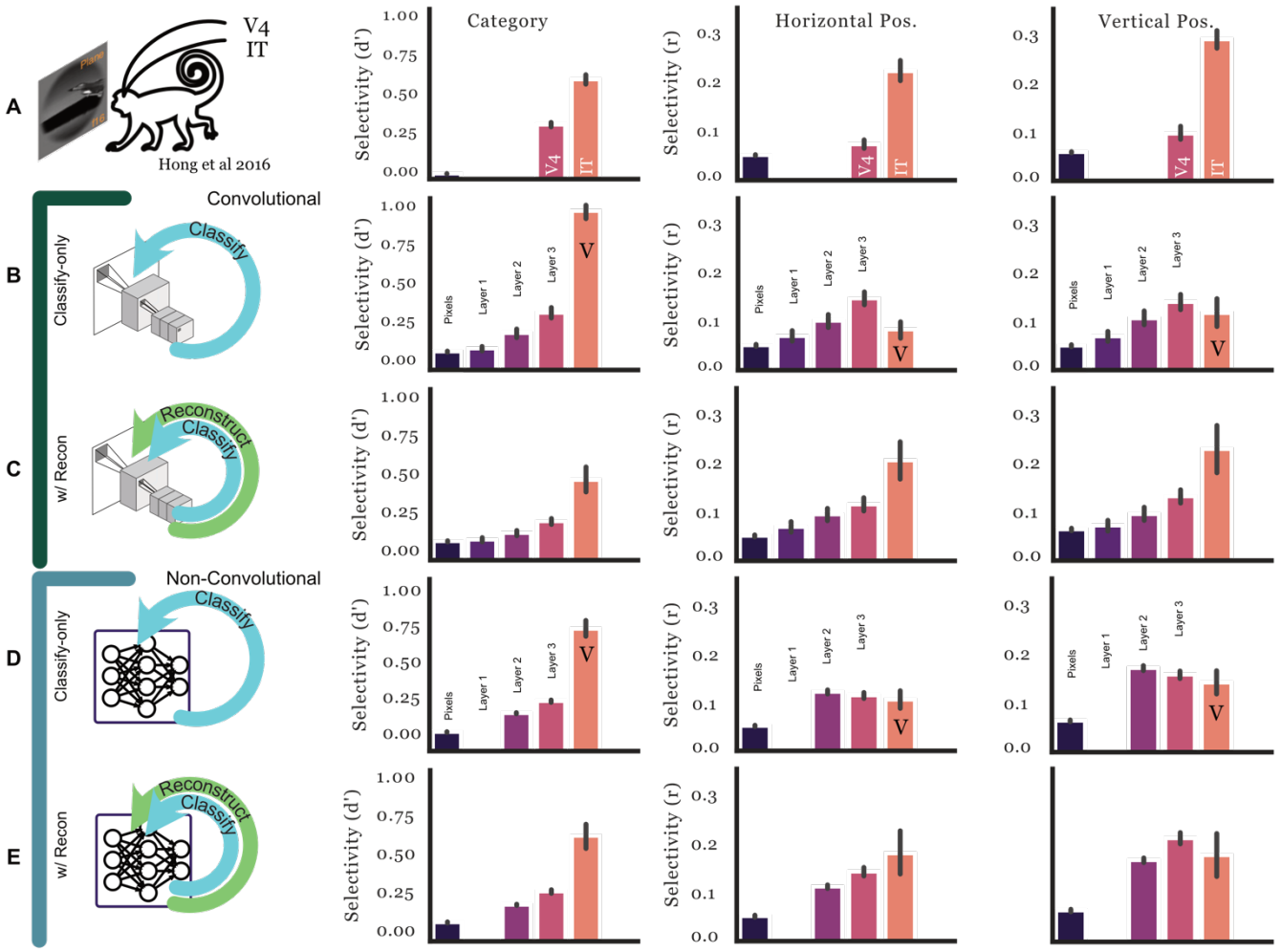
We would like to thank the DiCarlo lab for sharing their primate electrophysiology recordings with us. Special thanks to Alon Poleg-Polsky for thoughtful discussion and direction and to Doug Crawford, Shaiyan Keshvari, Martin Schrimpf, Rachel Sewell, and Heidi Sjoberg for helpful feedback on the manuscript. EC acknowledges funding by an NDSEG fellowship through the US Department of Defense. JZ acknowledges funding from CIFAR, the A.P. Sloan Foundation, Google, the Canada Research Chairs Program, and the Natural Sciences and Engineering Research Council of Canada (NSERC).

Fig. 1: Overview



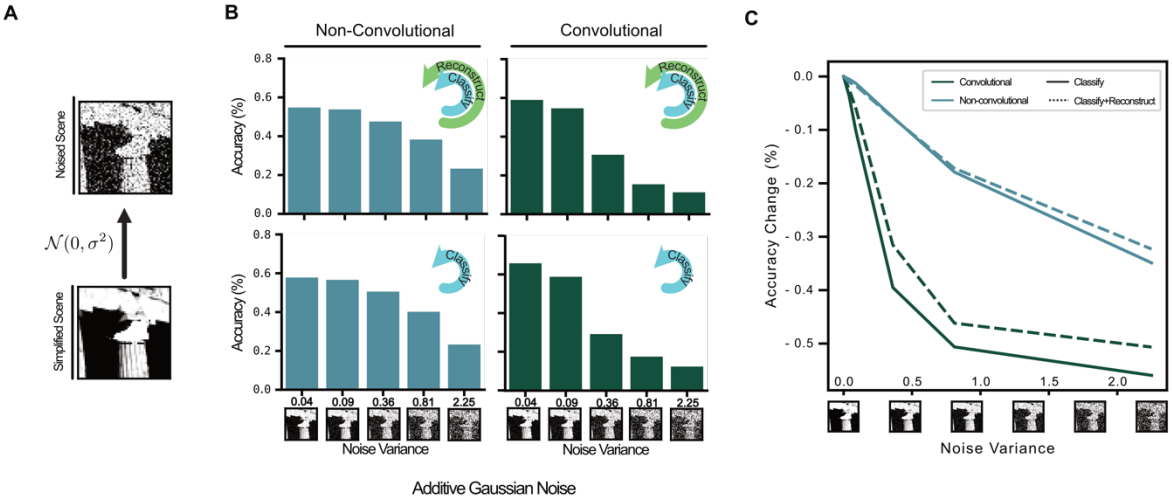
A) We constructed images of clothing items superimposed over natural image backgrounds at random eccentricities. **B)** We model the ventral stream as an encoder whose objective is to map input image (x) onto more abstract "latent" representations (*D* and *V*). In our models this entire latent space is represented by 70 artificial neurons (35 units in each of *D* and *V*) The generator network uses these latent representations (*D* and *V*) as input to reconstruct the object and its location within the scene. A separate linear decoder attempts to determine the object identity from the activities of the units in *V*. **C)** We trained both convolutional, and non-convolutional neural network architectures, on one of two tasks: object categorization ("classify"), or object categorization with concurrent image reconstruction. We note that, for the "pure" object recognition task, the generator network is superfluous. **D)** Neural networks with both architectures achieve comparable object recognition performance (accuracy) when using either classify-only and classify+reconstruct objective functions. This performance was assessed on held-out images, not used in training the networks.

Fig. 2: Comparisons of selectivity for visual scene properties



A) Category and position selectivity of single units recorded from macaque ventral stream (see Methods and Hong et al. 2016). **B&C)** Selectivity of units in the fully trained convolutional models optimized under classify-only objective (categorical cross-entropy) and the composite classify+reconstruct autoencoder objective. **D&E)** Non-convolutional or "all-to-all" models were also trained on both classify-only and classify+reconstruct. We measured property selectivity of both categorical and continuous valued category-orthogonal properties on units in the multi-electrode array data and each layer of the computational model encoders. We defined selectivity for categorical information on each unit in the dataset as the absolute value of that unit's discriminability (one-vs-all d-prime). We defined selectivity for continuous valued attributes (horizontal and vertical position) on each unit as the absolute value of the Pearson correlation coefficient. Unit activities for models were sampled using 10000 held out test images to generate activations at each layer of the model. For layers containing more than 192 units we randomly sampled 192 units for the analysis (to have a number of units similar to the number of IT units in the neural recordings).

Fig. 3: Noise generalization properties of models



A) Additive gaussian noise (mean=0) was used to corrupt 10,000 testing images at increasing levels. **B)** Each model (defined its architecture – convolutional or non-convolutional -- and the objective on which it was trained) was evaluated on images corrupted with increasing levels of gaussian noise. We show the accuracy at categorizing the objects in the noise-corrupted images. These images were from a held-out dataset, not used in training the neural networks. **C)** Convolutional neural networks are more sensitive to noise than are non-convolutional ones; they show a larger decrease in accuracy with increasing noise variance. Adding a reconstruction component to the network objective reduces this sensitivity. Similar results were obtained with a multiplicative noise model (Fig. S2), indicating that this result is not sensitive to the specific type of noise that corrupts the images.

Methods

Dataset and Augmentation

We constructed images of clothing items superimposed at random locations over natural image backgrounds. To achieve this goal, we used all 70,000 images from the Fashion MNIST dataset, a computer vision object recognition dataset comprised of images of clothing articles from 10 different categories. We augmented this dataset by expanding the background of the image two-fold (from 28x28 pixels to 56x56 pixels) and drawing dx and dy linear pixel displacements from a uniform distribution spanning 75% of the image field {-11,11}. Images were then shifted according the randomly drawn dx and dy values. After applying positional shifts, the objects were superimposed over random patches extracted from natural images from the BSDS500 natural image dataset to produce simplified natural scenes which contain categorical (1 of 10 clothing categories) and non-categorical (position shifts) variation. Random 56x56 pixel patches from the BSDS500 dataset were gray scaled before the shifted object images were added to the background patch (Fig 1A). All augmentation was performed on-line during training. That is, every position shift and natural image patch was drawn randomly every training batch instead of pre-computing shifts and backgrounds. This allows every training batch to be composed of unique examples from the dataset and prevents overfitting.

Primate Electrophysiology

Neural recordings were originally collected by the DiCarlo lab (Hong et al 2016) and shared with us for this analysis. In brief, neural recordings were collected from the visual cortex of two awake and behaving rhesus macaques using multi-electrode array electrophysiology recording systems (BlackRock Microsystems). Animals were presented with a series of images showing 64 distinct objects from 8 classes rendered at varying eccentricity in the animal's visual field. After spike-sorting and quality control this resulted in well-isolated single units from both IT (n=168) and V4 (n=128); higher-order areas in primate visual cortex. A full description of the data and experimental methods is given by Hong et al. (2016).

Computational models

Non-convolutional models were constructed by sequentially combining all-to-all (aka densely connected) layers. Any given layer uses the previous layers' output as input, multiplying the inputs (x) from by a weight matrix (w) and adds a bias to each unit in the output. Finally, this value is passed through a nonlinear activation function. Each layer outputs an activation vector of its units (y) which is function of its inputs (x).

$$y = \sigma\big((x \cdot w) + b\big)$$

The size of each layer in the models used in our experiments were chosen to have layer sizes with roughly similar proportions to the number of output neurons in corresponding brain areas of ventral stream (Felleman and Van Essen 1991).

Area	# output neurons (10^6)	# layer outputs	Computational model
		3136 (56 x 56)	Layer 0 (input image)
V1	37	3000	Layer 1
V2	29	2000	Layer 2
V4	15	2000	Layer 3
IT	10	70 (35+35)	Layer 4

Similar to the non-convolutional models, the convolutional models were constructed by sequentially combining convolutional layers. Each convolutional layer receives as input a spatially arranged map from the prior layer. A filter kernel is multiplied against the input at each spatial location in the input, and the resultant value is added to the bias and passed through the nonlinear activation function.

The convolutional models described in our paper were constructed according to the table below:

	Output Size	*Kernel Size*	*Activation Function*	*Dropout rate*	*Batch Normalization Momentum*
Input	56 x 56	N/A	N/A	N/A	N/A
Layer 1	28x28x16	3x3	LeakyReLU	25%	0.8
Layer 2	14x14x32	3x3	LeakyReLU	25%	0.8
Layer 3	7x7x64	3x3	LeakyReLU	25%	0.8
Layer 4	70(35+35)		Linear	0%	0.8

Models using the composite classify-reconstruct objective (see below) need an additional generator network to reconstruct the original stimulus input from the latent representation. The generator network (G) uses a residual convolutional neural network (ResNet) which has achieved state of the art performance in natural image generation.

The generator network uses is comprised of Deconvolutional layers and its architectural hyperparameters directly mirror those in the convolutional encoder.

Objective functions and training parameters

Models optimized for classification use categorical cross-entropy for the objective function. Categorical cross-entropy (XENT) is a commonly used objective function in machine learning to train neural network classifiers. Multilabel cross-entropy is calculated according to the equation below where M is the total number of classes

$$\boldsymbol{XENT} = -\sum_{c=1}^{M} \boldsymbol{y_c} \cdot \boldsymbol{ln}(\hat{\boldsymbol{y}}_{\boldsymbol{c}})$$

Here, $\boldsymbol{y_c}$ is the true category label, represented as a one-hot vector, and $\hat{\boldsymbol{y}}_{\boldsymbol{c}}$ is the network output obtained from the linear readout of population V (see Fig. 1).

Models with an objective function term for reconstructing the original input scene use pixel-wise sum of squared error (SSE) between the input and the generator's output ($\hat{\boldsymbol{x}}$).

$$\boldsymbol{SSE} = \sum (\boldsymbol{x} - \hat{\boldsymbol{x}})^2$$

We trained each model in our experiment until classification accuracy plateaued on a validation dataset of 512 objects from the 10,000 test images in the fashion MNIST dataset.

Model Evaluation

After training performance plateaus, 192 randomly chosen unit activations from Layers 1-3 in the encoder model (Fig 1B) were used in comparisons with primate ventral stream electrophysiology. Unit activations were generated using a random sample from held out test images (not used during training). As in a (simulated) electrophysiology experiment, each image was input to the network, and the corresponding unit activations were recorded. We then analyzed these unit activations in the same way as we did the firing rates recorded in monkey visual cortex.

We measured selectivity of our artificial neurons in the same way as Hong et al 2016 (they call these measures "performance" instead of selectivity). For continuous-valued scene attributes (e.g. horizontal position) we measured selectivity as the absolute value of the Pearson correlation between the neuron's response and that attribute in the stimulus image. For categorical properties (e.g. object class) we measure selectivity as the one-vs-all discriminability (d').

We quantified the similarity of each models' layer-wise selectivity to corresponding layers in primate ventral stream using Fisher's Combined Probability Test (FCT). As discussed in the main paper, we first used the Welch's unpaired t-test to calculate p-values model-VS pairs for all selectivity metrics in the corresponding layers, then used the FCT to combine those p-values into a single likelihood measure that reflects the likelihood of observing the monkey physiology data, under the hypothesis that those data are drawn from the same distribution as the units computational model: a larger p-value corresponds to a model that more closely matches the monkey data.

Noise Generalization

To test how well each model performs under increasing amounts of noise, we added noise generated from two types of distributions: uniform-multiplicative and gaussian-additive. Both distributions had mean 0 and variance values drawn from the following table:

Uniform:	0.01	0.09	0.36	0.81	2.25
Gaussian:	0.2	0.3	0.6	0.9	1.5

Gaussian noise was added the input image after adding natural background and position shift. Uniform noise was applied multiplicatively according to the following formula where I is the original image and η is a 56x56 matrix of values drawn from a uniform distribution with mean=0 and one of the variance values above:

$$I = I + (\eta * I)$$

References

1. Felleman, D.J. & Van Essen, D.C. *Cereb. Cortex* **1**, 1–47 (1991).
2. Yamins, D.L.K. & DiCarlo, J.J. *Nat. Neurosci.* **19**, 356–365 (2016).
3. Yamins, D.L.K. et al. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
4. Cadieu, C.F. et al. *PLoS Comput. Biol.* **10**, e1003963 (2014).
5. Güçlü, U. & van Gerven, M.A.J. *J. Neurosci.* **35**, 10005–10014 (2015).
6. Stokes, M., Thompson, R., Cusack, R. & Duncan, J. *J. Neurosci.* **29**, 1565–1572 (2009).
7. O'Craven, K.M. & Kanwisher, N. *J Cogn Neurosci* **12**, 1013–1023 (2000).
8. Chen, Y. & Crawford, J.D. *Annals of the New York Academy of Sciences* **46**, 774 (2019).
9. Hong, H., Yamins, D.L.K., Majaj, N.J. & DiCarlo, J.J. *Nat. Neurosci.* **19**, 613–622 (2016).
10. Richards, B.A. et al. *Nat. Neurosci.* **22**, 1761–1770 (2019).
11. Carandini, M. & Heeger, D.J. *Nat Rev Neurosci* **13**, 51–62 (2011).
12. Majaj, N.J., Hong, H., Solomon, E.A. & DiCarlo, J.J. *J. Neurosci.* **35**, 13402–13418 (2015).
13. Li, Q., Hu, J., Ding, J. & Zheng, G. *Biostatistics* **15**, 284–295 (2014).
14. Geirhos, R. et al. (2018).
15. Freud, E., Plaut, D.C. & Behrmann, M. *Trends in Cognitive Sciences* **20**, 773–784 (2016).
16. Sereno, A.B. & Lehky, S.R. *Front Comput Neurosci* **4**, 159 (2011).
17. Rezai, O., Stoffl, L. & Tripp, B. *Neural Netw* **121**, 122–131 (2020).