Models of the ventral stream that categorize and visualize images

Elijah Christensen[1], Joel Zylberberg[2,3]

[1]Department of Physiology and Biophysics, University of Colorado Anschutz Medical Campus
[2]Learning in Machines and Brains Program, CIFAR, Toronto, ON Canada
[3]Centre for Vision Research, and Department of Physics and Astronomy, York University, 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. These findings raise the question of whether primate VS may do more than "just" object recognition. 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 visual stimuli, than do models that "just" identify objects. A thorough understanding of VS computations might thus require considerations beyond object recognition. Scene reconstruction is a candidate for such consideration.

Introduction

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 previously encountered and novel scenes[6,7], suggesting that the VS can *generate* visual scenes in addition to identifying objects within scenes. 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 the notion that the 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 responses 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. Finally, we used canonical correlation analysis (CCA) to directly compare the representations of input images within the neural networks, to those of the monkey. 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 do those optimized for task (a). This suggests that a full understanding of visual ventral stream computations might require considerations other than object recognition, and that scene reconstruction is a promising candidate for the "other" computations occurring within the VS.

Results

Computational Models

To identify the degree to which different computational objectives describe ventral stream physiology, we optimized deep convolutional neural network (CNN) 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 orientation, and placed over a natural image background (Fig. 1A).

The models each had a total of four layers of processing between their inputs and these latent representations; we believe these latent representations correspond to area IT, for reasons we discuss below. The four layers of processing that precede the latent representation in our model are then (loosely) analogous to lateral geniculate nucleus (LGN), V1, V2, and V4 of the mammalian visual system. The visual inputs to the model had normalized luminance values, mimicking the normalization observed at LGN[11]. The connectivity between neurons in each layer (and the artificial neurons' biases) were optimized within each model, 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 network to reconstruct the object from the latent representation. (See Methods for details about the optimization procedure).

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. Reasonable performance on the categorization task was obtained in both 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 of objects superimposed over natural image backgrounds, at different locations and orientations. Many neurons in each area were simultaneously observed in these experiments. 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 object category and position increased along the ventral stream (Fig 2A), as reported by Hong et al[9]. This indicates that both categorical and non-categorical 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): both category and position are represented more explicitly with each subsequent layer of the model. 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).

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 input into our models the same images that were shown to the monkeys in the physiology experiments. We then extracted the activations of the artificial neurons at each layer of our computational models, and we used Canonical Correlation Analysis (CCA)[17] to compare those artificial neurons' activations to those recorded in monkey V4 and IT (See Methods). In brief, CCA assesses the degree to which weighted sums of our neural network unit activations correlate with weighted sum of the neuron firing rates observed in the monkey experiments. It can thus test for similarity in how the images are represented by the neural networks, and the monkey, without requiring us to assign each neural network unit to a specific neuron in the monkey experiments. Similar to regular correlation analysis, CCA correlations of 0 indicate no relation between the neural network and monkey visual representations, while a value of 1 indicate perfect similarity.

This analysis indicates that, both monkey V4 and IT are a better match to the deeper layers in the network trained on the composite task. (I.e., the canonical correlation values increase consistently with depth for the "full recon" model in Fig. 3). For the model trained on the pure object recognition task, the canonical correlation values rapidly saturate as one considers progressively deeper layers in the model (see "no recon" data in Fig. 3). Overall, the best match to the V4 and IT data is found in the deepest layer (the latent representation) of the network trained on the composite task.

Discussion

Here we report evidence that convolutional neural networks (CNNs) 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. Networks trained on the composite task also show better quantitative matches to the image-evoked firing patterns recorded in primate VS. These results are unexpected, as prior work posits that networks optimized strictly for object recognition should form the best models of primate VS[2,4,9,10] because each

network's similarity to primate VS recordings correlates strongly with that network's object recognition accuracy. Our network trained on the composite task, however, provided a better match to the primate VS data (Fig. 3) despite having slightly lower object recognition performance (Fig. 1D) than the network trained purely for object categorization. Our results suggest that the image representations in the primate ventral stream could require a different functional explanation (not simply object recognition).

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 we studied promotes retention of both category and category-orthogonal information because both are necessary to reconstruct the stimulus. 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.

Materials and Methods

Dataset and Augmentation

Our goal was to study the object representations, scene reconstruction, and representation of non-categorical information, within artificial neural networks. To achieve that goal, we trained the neural networks to take in images, and either categorize the objects within them, or categorize the objects *and* reconstruct the input (e.g., a semi-supervised autoencoder[18]). To train these networks, we required images that varied in categorical, and in non-categorical, properties. For that reason, 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, and rotated according to randomly drawn angles between -30 and +30 degrees. After applying positional and rotational 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 and rotation 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, rotation 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 combinations of objects, backgrounds, rotations, and shifts, helping to prevent overfitting. (E.g., note that the validation accuracy in Fig. 1D increases monotonically during training, instead of decreasing at later times as would be expected if the models were overtraining).

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 position in the animal's visual field, and with variation rotations. 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

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)
LGN	10	3000	Layer 1
V1	37	3000	Layer 2
V2	29	2000	Layer 3
V4	15	1000	Layer 4
IT	10	500	Latent

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	112 x 112	N/A	N/A	N/A	N/A
Layer 1	56x56x16	3x3	LeakyReLU	25%	0.8
Layer 2	28x28x32	3x3	LeakyReLU	25%	0.8
Layer 3	14x14x64	3x3	LeakyReLU	25%	0.8
Layer 4	7x7x128	3x3	LeakyReLU	25%	0.8
Latent	500		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. Our models can be found on Github (https://github.com/elijahc/vae).

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

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

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

Model Evaluation

Feature Selectivity (Fig. 2):

After training performance plateaus, 5-fold sampling of 250 randomly chosen unit activations from Layers 1-4 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, described below.

First, we measured selectivity of our artificial neurons to different image attributes, 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').

Canonical Correlation Analysis (Fig. 3):

Next, we quantified the similarity of each models' layer-wise selectivity to corresponding layers in primate ventral stream using Canonical Correlation Analysis (CCA)[17]. CCA finds a set of weights used to project both the primate electrophysiology results and our own model unit activations into a lower dimensional space and measures the correlation of the projections in this space. The projection weights are optimized to maximize correlation in the lower dimension. We use 10 projection dimension for this analysis and report the average over the (optimized) correlations of those 10 dimensions. In analogy to the monkey experiments, we performed these analyses on randomly-chosen sets of 250 units from our models; this matches the pseudo-random sampling of neurons with the implanted electrode arrays, and approximates the number of neurons recorded in each experiment. We repeated the analysis for ten different random draws of unit activations and report the distribution of correlations over those ten draws (Fig 3).

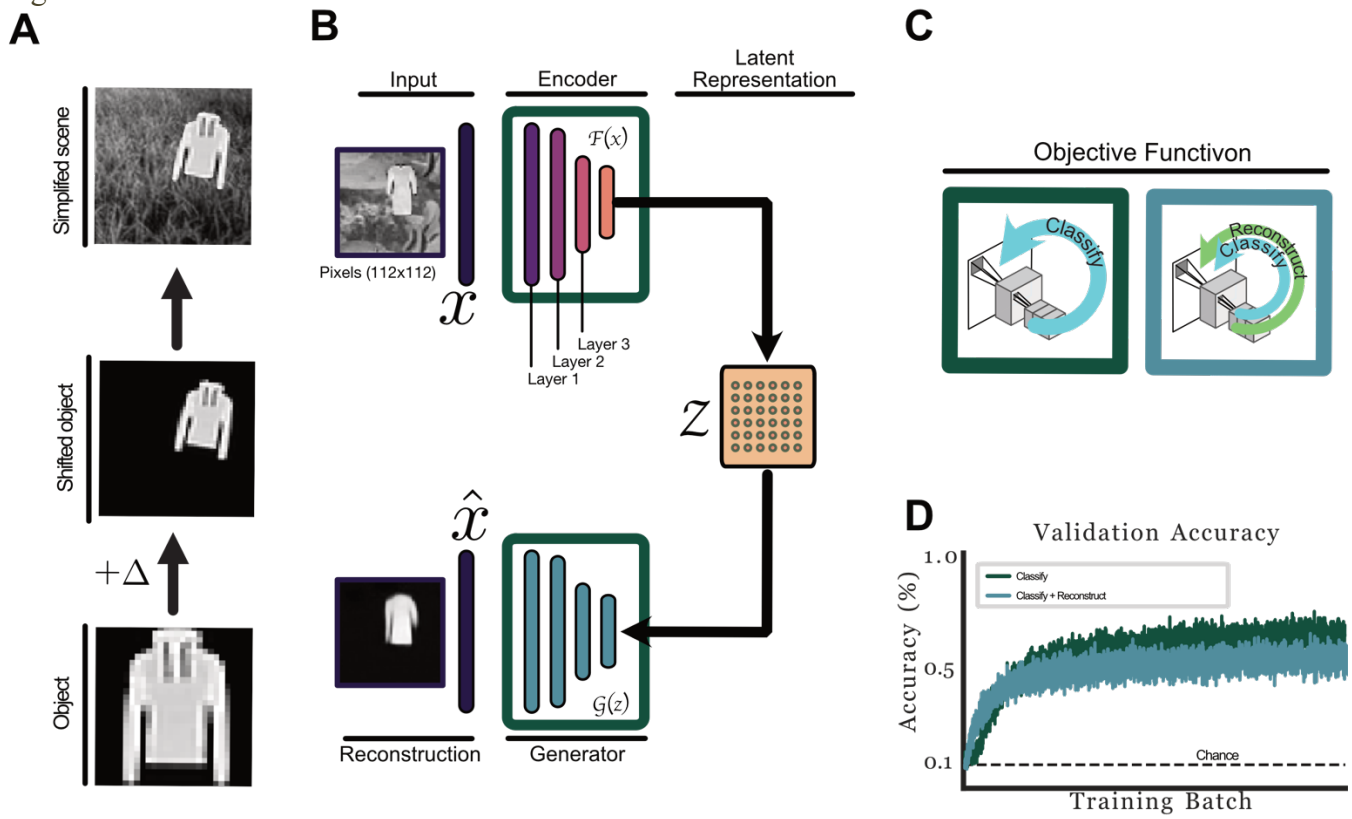
Acknowledgements

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).

References

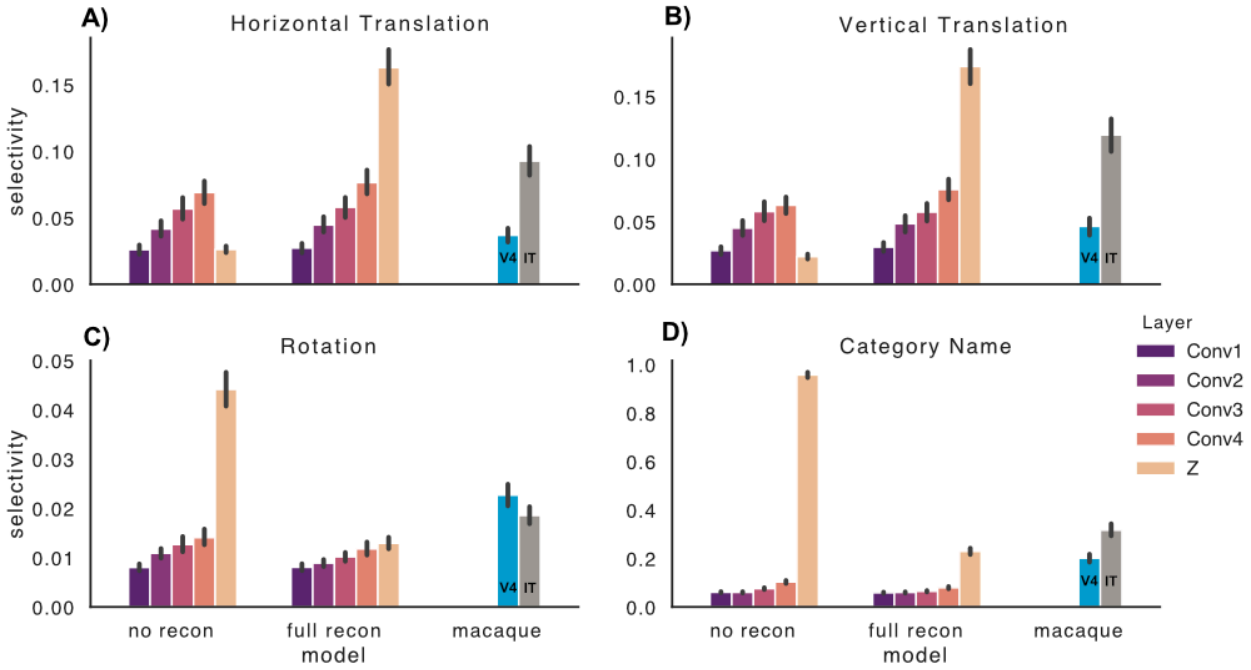
1. Felleman, D.J. & Van Essen, D.C. *Cereb. Cortex* **1**, 1–47 (1991).
2. Yamins, D.L.K. & DiCarlo, J.J. *Nat. Neurosci.* **19**, 356–365 (2016).
3. Yamins, D.L.K. et al. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
4. Cadieu, C.F. et al. *PLoS Comput. Biol.* **10**, e1003963 (2014).
5. 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. Sussillo, D., Churchland, M. M., Kaufman M. T., & Shenoy, K.V. *Nature neuroscience*, 18(7):1025–1033 (2015)
18. Cheung, B., et al. “Discovering hidden factors of variation in deep networks.” *arXiv preprint*, arXiv: 1412.6583 (2014)

Fig. 1: Overview



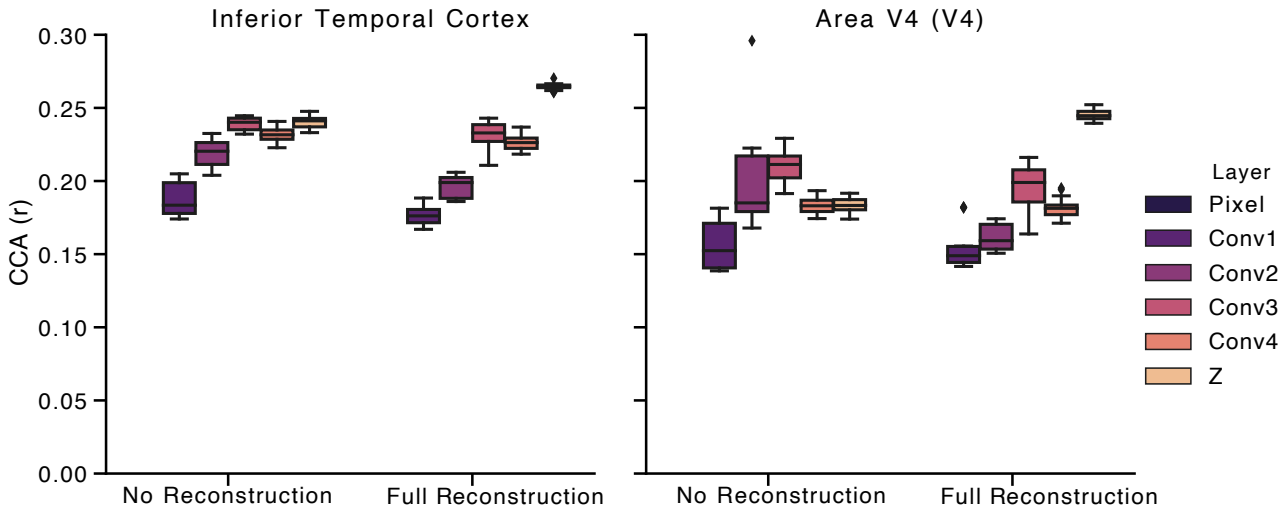
A) We constructed images of clothing items superimposed over natural image backgrounds at random eccentricities. **B)** We model the ventral stream as an encoder whose objective is to map input image (x) onto more abstract "latent" representations (z). In our models this latent space contains 500 artificial neurons. The generator network ($\mathcal{G}$) uses these latent representations (z) 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 z. **C)** We trained convolutional neural networks 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 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



Selectivity of units in the fully trained convolutional models optimized under classify-only objective (categorical cross-entropy) and the composite classify+reconstruct semisupervised autoencoder objective[18]. We measured property selectivity of both categorical (**D**) and continuous valued category-orthogonal properties (**A**, **B**, **C**) on units in the multi-electrode array data from Hong et al. (2016), and from units in 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. We randomly sampled 250 units from each layer of each model for the analysis. Error bars represent the S.E.M. over the observed set of units.

Fig. 3: Canonical Correlation Analysis



We used Canonical Correlation Analysis (CCA) to quantify how similar the responses in the layers of each model were to primate electrophysiology data in both inferior temporal cortex (IT) and visual area V4 (V4). We used random draws of 250 unit activations in each layer of the fully trained convolutional models optimized under the classify-only objective (categorical cross-entropy, left in each panel) and the composite classify+reconstruct semisupervised autoencoder objective. We repeated this process for 10 random draws of the neural network unit activations, and display the distribution of the resultant CCA correlation values (over those 10 draws) as a box and whisker plot. Lines within the filled bar indicate the mean, and filled rectangle corresponds to the interquartile range.