Convolutional Neural Networks Mimic the Hierarchy of Visual Representations in the Human Brain

Pulkit Agrawal
UC Berkeley
pulkitag@berkeley.edu

Dustin Stansbury
UC Berkeley
stan_s_bury@berkeley.edu

Jitendra Malik
UC Berkeley
malik@eecs.berkeley.edu

Jack Gallant
UC Berkeley
gallant@berkeley.edu

Abstract

The human brain transforms visual inputs into information that is useful for semantic tasks like object recognition and scene interpretation. How the brain performs this transformation is an open question in neuroscience. Recently, Convolutional Neural Networks (ConvNets) have been successfully used for transforming image pixels into features useful for object recognition. Just like the early and late stages of visual processing in the brain, the lower and higher layers of a ConvNet represent gabor like and semantically meaningful features respectively. Based on this, we hypothesized that intermediate layers of ConvNets and the human brain may use similar features for representing visual information. Using fMRI recordings of human subjects viewing natural images, we show that the hierarchy of visual representations in a ConvNet trained for object recognition mimics the hierarchy of visual representations in the human brain. This result suggests that understanding visual representations in the ConvNet can help us understand the visual representations in the human brain.

1 Introduction

How does the human brain transform visual information captured by the retina into information useful for semantic tasks like object recognition and scene interpretation? We know that this transformation is performed by a hierarchically organized system of visual areas within the visual cortex. However, we do not know how the brain computes this transformation. Past studies have found that visual regions of interests (ROIs [1]) like V1, V2 located in posterior visual cortex appear to represent low-level visual features such as oriented edges [2] , gabors [3] and local motion-energy features [4]. Visual ROIs like Fusiform Face Area (FFA [5, 6]), Extrastriate Body Area (EBA [7]) and Parahippocampal Place Area (PPA[8]) located in anterior visual cortex appear to represent high-level semantically meaningful features useful for detecting faces, bodies and understanding visual scenes. Further, it is believed that visual ROIs like V4 located in intermediate visual cortex represent mid-level visual features that are useful grouping, figure-ground segmentation and representing contours [9]. However, the visual representations and computations performed in intermediate visual ROIs is poorly understood.

One way to address the question of understanding how the brain transforms low-level visual representations into high-level visual representation is to build a computational model that takes images/videos as inputs and outputs an accurate prediction of brain activity across the visual cortex. An accurate prediction of brain activity would imply that the constructed model and the brain represent visual information using similar features. The similarity of features by itself would be insufficient

to make conclusions about the exact computations performed by the brain. However, such a finding would suggest that the constructed model is a plausible computational hypothesis for how the brain transforms low-level features into high-level features. Further, such a model could also be used to investigate the nature of visual representations in different parts of the visual cortex.

Given that brain is a complex non-linear processing system, it is unlikely that an analytical solution to the problem of constructing such a model would exist. Past studies have addressed this concern by breaking down the process of predicting brain activity elicited in response to stimulus images into two steps. In the first step, a feature space that provides a linearizing transformation between the stimulus images and measured brain activity is constructed. In the second step, regularized linear regression is used to find a set of weights that predict brain activity from the feature representation of images. This framework for predicting brain activity has been called the encoding model approach [10, 11, 12, 13, 14, 3]. Past studies used manually constructed feature spaces for predicting brain activity. For instance, [3, 4] predicted brain activity in visual ROIs like V1, V2 using Gabor features. [13, 12] used linguistically constructed feature spaces that indicated the presence or absence of multiple object categories in images. These studies were only only able to predict brain activity in anterior visual cortex (i.e. ROIs like FFA, EBA, PPA). Moreover, these studies were unsatisfying because they did not provide any explanation for how the brain converts stimulus images into semantically meaningful information. To date, there exists no model that can predict brain activity throughout the visual cortex starting from image pixels.

Instead of manually defining features, an alternative is to use machine learning techniques to learn features that are optimal for predicting brain activity. However, it is unlikely that these techniques would work because non-linear machine learning methods require large amounts of training data and brain activity recordings are not available in plenty. This is because, collecting brain activity data is both a tedious and a costly process. Another way to learn features is by training models for performing the same tasks that the human visual system performs. After all, it is reasonable to assume that visual processing and representations in the brain are optimized for the visual tasks it must perform. Moreover, large amounts of data are publically available for training models for performing tasks like object recognition that are also performed by humans [15].

Recently in the field of computer vision, a class of computational models called as Convolutional Neural Networks (ConvNets [16]) have been found to be very successful on the task of object recognition [17]. Multiple considerations suggest that the visual features of a ConvNet are a good candidate for studying visual features represented by the brain. Firstly, the brain and the ConvNet are both adept at the common task of object recognition. Secondly, the brain and the ConvNet both represent visual information hierarchically. For instance, the ConvNet architecture proposed by [17] represented images by a seven-layered hierarchy of visual features. Lastly, some past studies have shown that the lower layers of the ConvNet feature hierarchy represent visual features such as edges and corners whereas the higher layers represent visual features that are more useful for object recognition [18, 19]. These three facts taken together suggest that low and high-level visual features represented by the brain and the ConvNet are likely to be similar. If it is the case that low and high-level features represented by the brain and the ConvNet are similar, it is likely that mid-level features represented by the brain and the ConvNet are also similar.

In this work we tested the above hypothesis by investigating the relationship between the hierarchies of visual representations in the human brain and a ConvNet trained for the task of object recognition. The method and results of our investigation are presented in section 2 and section 3 respectively. We provides a discussion of the implication of the results and a comparison with related previous work in section 4. The conclusions of our study are mentioned in section 5.

2 Method

For studying the relationship between the visual representations of the ConvNet and the human brain we constructed computational models for predicting brain activity from visual representations of the ConvNet (see figure 1). First, we trained a seven layered ConvNet with the architecture proposed by [17] for the task of classifying 1.2M million natural images into 1000 distinct object categories (ILSVRC-2012 challenge [15]) using the publically available software [20]. In the remainder of this paper, the term ConvNet refers to this particular ConvNet. This ConvNet transformed input images

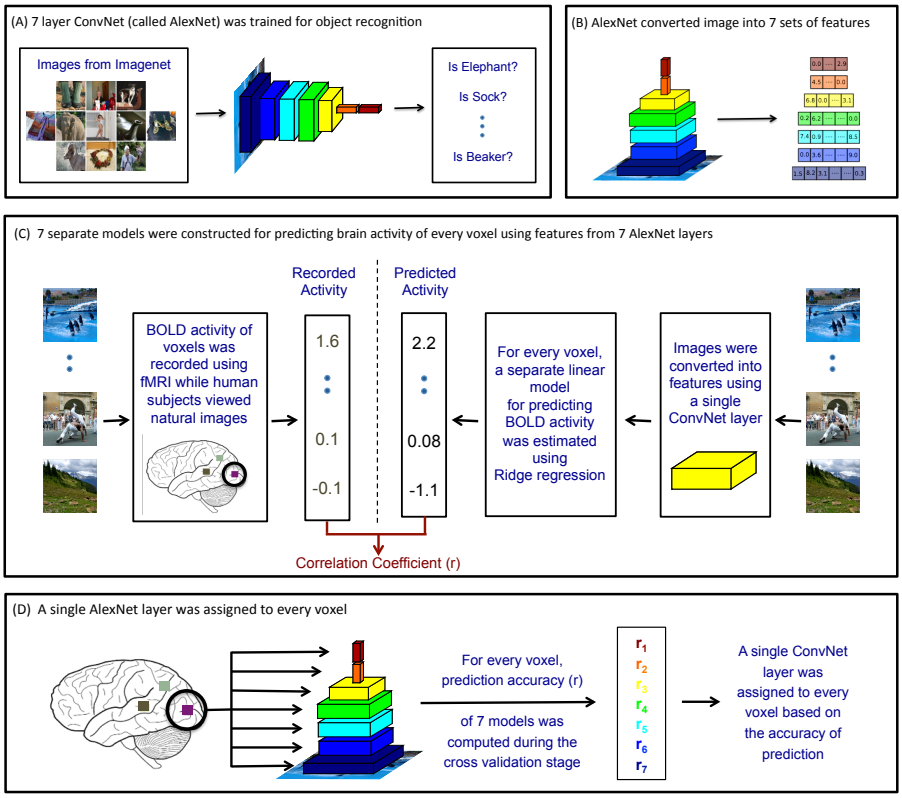

Figure 1: Description of the method for predicting brain activity using ConvNet features. First, a seven layered ConvNet was trained for the task of recognizing 1000 distinct object categories using a collection of 1.2M labelled images (panel A) [17, 15]. This ConvNet extracted seven sets of features (from seven layers) for a given input image (panel B). ConvNet features were used to predict brain activity (i.e. BOLD activity measured using fMRI) of four human subjects while they passively viewed a separate collection of natural images (panel C). For every voxel, seven separate models for predicting BOLD activity were constructed using features extracted from seven layers of the ConvNet. Based on the accuracy of prediction (measured as correlation coefficient), an optimal ConvNet layer was assigned to every voxel (panel D).

into seven set of features (one from each of the seven layers). These features were used to predict brain activity.

The brain activity data for this study were functional magnetic resonance imaging (fMRI [21]) recordings of human brain activity (specifically, the blood-oxygenation-level-dependent (BOLD) signal), recorded continuously while four subjects passively viewed a series of static photos of color natural scenes. These subjects have been referred to as S1, S2, S3 and S4 in the remainder of the paper. We used the fMRI data that was previously used by [12]. This study measured brain activity elicited by 1260 images shown twice each (train set of images), and another set of 126 images shown 12 times each (test set of images). Activity was measured in approximately 100,000 voxels (i.e., volumetric pixels) located in the cerebral cortex of each subject. We followed the same procedure for pre-processing the fMRI data as outlined in [12].

2.1 Constructing Models for Predicting Brain Activity

For every voxel, a separate model was constructed for predicting its BOLD activity from the given feature representation of the image. Ridge regression was used to find a set of weights that predicted voxel's BOLD activity using the training set of 1260 images. A single regularization parameter was chosen for all voxels, using five-fold cross-validation [13]. The accuracy of each model for each voxel was expressed as the correlation (r) between predicted and recorded voxel activity in response

to images in the test set. The explained variance in each voxel's responses was calculated as the square of correlation coefficient (r2) [10]. Prediction accuracy was deemed statistically significant if the correlation had a p-value < 0.001 (see supplementary materials for more details).

The modelling framework described above implicitly assumes that voxel responses are stationary (i.e. the BOLD activity of a voxel is only a function of the input image and will be the same every time the same image is presented as stimulus). However, the voxel responses can be non-stationary due to either the inherent non-stationarity in firing of individual neurons that constitute the voxel or due to noise in fMRI measurements. As our modelling framework is incapable of dealing with non-stationarity, we only fit models to voxels that are approximately stationary. The stationarity of a voxel can be estimated by calculating the repeatability in BOLD activity of the voxel expressed as the Signal to Noise Ratio (SNR). The method for computing the SNR is detailed in the supplementary materials. In this work, SNR has been expressed in terms of p-values (p_{SNR}). Note that this p_{SNR} is different measure than the p-value of the prediction accuracy. In this work, we have only considered voxels with $p_{SNR} < 0.001$.

2.1.1 Convolutional Neural Network (ConvNet) Model

After the ConvNet was trained for object recognition, it was used to transform all images used in the fMRI study into seven sets of features. The images used in the fMRI study were seperate from images used for training the ConvNet. Each set of feature corresponded to the feature representation of images produced by a single layer of the ConvNet. The first five layers of the ConvNet performed convolutions (denoted conv-1 through conv-5) and the last two layers were fully connected (fc-6, fc-7) (see supplementary materials for more details). For every voxel, seven separate sets of weights were estimated for predicting brain activity from these seven feature spaces. An optimal ConvNet layer was determined for every voxel based on the prediction accuracy of voxel activity measured during the cross-validation stage (see figure 1).

2.1.2 Baseline Model

In order to compare the prediction accuracy of the ConvNet with previously published models, a baseline model was constructed by combining the Gabor Wavelet model (GW; [3]) and the 19-Category model (19-Cat; [11, 12]). The GW and 19-Cat model have been shown to accurately predict brain activity in early (V1, V2) and late (PPA, FFA, EBA, OPA) visual areas respectively. More details on these two models has been provided in the supplementary materials. The Baseline model was constructed in the following way: For every voxel two sets of weights were independently estimated for predicting BOLD activity from GW and 19-Cat features. Each voxel was then assigned either to the GW or the 19-Cat model. This assignment was made based on the accuracy of the GW and the 19Cat models in predicting BOLD activity of the voxel measured during the cross validation stage. The model obtained after this assignment has been called the Baseline model.

3 Results

3.1 ConvNet predicts brain activity across the visual cortex

Any hypothesized feature space provides useful insights into brain representations only to the extent that it accurately predicts brain activity elicited under naturalistic conditions. The prediction accuracy of the ConvNet model was evaluated using the test set of 126 images and evoked BOLD activity (that were separate from set of images used for fitting the model). Figure 2 shows prediction accuracy of the ConvNet model fit to voxels distributed across visual cortex for subject S1. From this figure it can be concluded that the ConvNet model transforms image pixels into features that make significant predictions (i.e. p-value < 0.001) of BOLD activity across the visual cortex.

If it is the case that ConvNet model can provide insights into visual representations in the human brain beyond what is already known, then the ConvNet model must predict brain activity with accuracy higher than previously published models. We compared the accuracy of ConvNet model with the Baseline model across multiple visual ROIs using the following two metrics: The percentage of significantly predicted voxels in a ROI and the percentage explained variance in the BOLD response within a ROI.

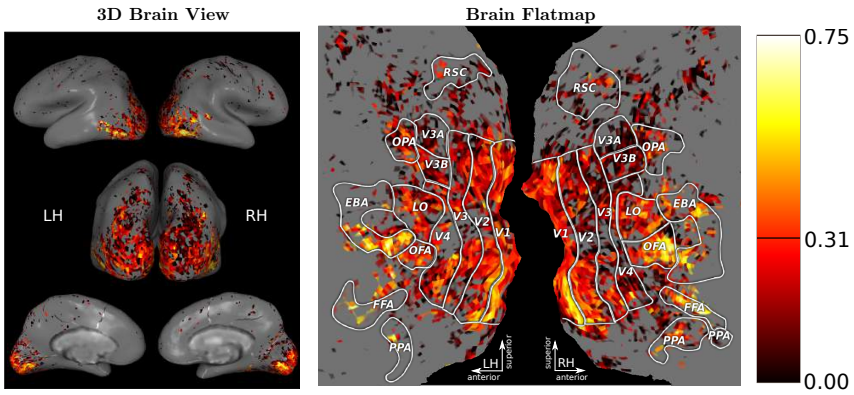

Figure 2: Accuracy of the ConvNet model in predicting brain activity of subject S1. The accuracy was measured as the correlation coefficient (Pearsons r) between the predicted and recorded brain activity in response to the test set of images. The color of each voxel reflects the prediction accuracy of the ConvNet model. Hotter colors reflect higher accuracy of prediction. The statistical significance of each voxel was evaluated individually and the mean cutoff value of r for the voxels with p-value < 0.001 was found to be 0.306 ± 0.008. The voxels with low SNR (i.e. p_{SNR}-value > 0.001) have been shown in gray.

Table 1: Comparing the accuracy of the ConvNet with the Baseline model for predicting brain activity across several visual ROIs. The accuracy was quantified using two metrics - the percentage of significantly predicted voxels (% Significant) and the mean explained variance (expressed as percentage, % Variance) in the BOLD activity of each ROI. The table reports the mean $\pm$ standard deviation of these metrics computed using 1000 bootstrap runs (see supplementary materials for details). The ConvNet model is as good or better than the baseline model in almost all ROIs and outperforms the baseline model in intermediate visual ROIs like V4, LO and OFA.

Measure	**Model**	**ROI**								
		V1	V2	V3	V4	LO	OFA	FFA	EBA	PPA
% Significant	ConvNet	32.8 ± 2.9	24.6 ± 1.9	16.3 ± 1.6	17.7 ± 1.6	41.7 ± 2.2	67.3 ± 4.3	69.2 ± 3.1	60.1 ± 3.4	47.7 ± 2.0
	Baseline	32.6 ± 2.9	26.6 ± 2.6	13.9 ± 1.9	11.0 ± 1.5	32.4 ± 1.9	53.3 ± 3.6	65.0 ± 3.5	57.6 ± 3.3	47.7 ± 2.6
% Variance	ConvNet	8.1 ± 0.6	6.4 ± 0.4	4.5 ± 0.3	4.7 ± 0.3	10.8 ± 0.8	17.4 ± 2.0	19.7 ± 2.2	15.5 ± 1.6	14.8 ± 1.3
	Baseline	7.5 ± 0.6	6.4 ± 0.5	4.0 ± 0.3	3.5 ± 0.2	8.4 ± 0.6	13.8 ± 1.3	17.8 ± 1.7	14.6 ± 1.4	14.2 ± 1.4

For making this comparison, voxels belonging to the same ROI were grouped across all the subjects. The percentage of significantly predicted voxels was calculated as the percentage of voxels within a ROI for which BOLD responses were predicted with p-value < 0.001. The explained variance in BOLD response of each ROI was calculated as the mean explained variance in BOLD responses of voxels assigned to the ROI (see supplementary materials for more details). The results are reported in table 1 indicate that ConvNet and the Baseline model make comparable predictions in early and late visual areas. The ConvNet model outperforms the Baseline model in intermediate visual areas. This suggests that the ConvNet model might provide novel insights about visual features represented by the intermediate visual cortex.

3.2 The hierarchy of visual representations in the ConvNet mimics the hierarchy of visual representations in the human brain

Does the ConvNet model provide insights into how the brain transforms low-level visual features into high-level visual features? If it is the case that the ConvNet provides a plausible computational hypothesis for how the brain transforms low-level visual features into high-level visual features then the low, mid and high-level features represented in both the systems must match.

To investigate if this was the case, we plotted the ConvNet layer assigned to every voxel on a flatmap of the brain (figure 3). Each voxel in this figure has been color-coded to reflect its corresponding ConvNet layer. The figure shows that lower, middle and higher layers of the ConvNet were optimal for predicting BOLD responses in posterior, intermediate and anterior parts of the visual cortex. This implies that low, mid and high-level visual features represented by the ConvNet are related

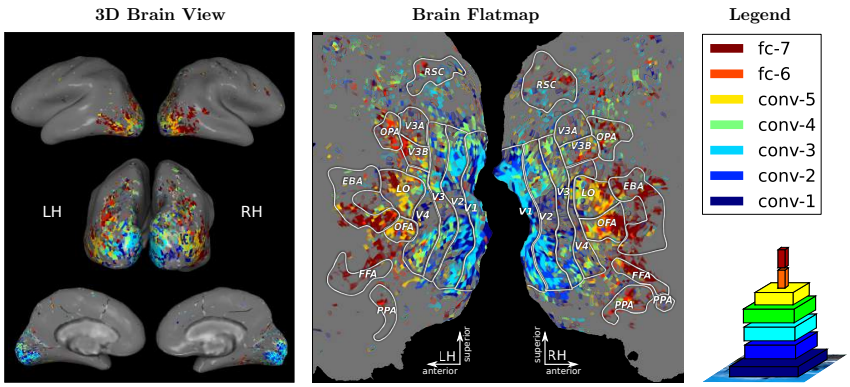

Figure 3: Relationship between the feature hierarchies of ConvNet and the human brain. Voxels have been color coded to show the ConvNet layer that was found to be optimal for predicting their activity. The voxels with p_{SNR}-value > 0.001 are shown in gray. The alpha channel of the voxel colors has been modulated to reflect the accuracy of the ConvNet model in predicting the BOLD activity of voxels. The alpha value for all voxels predicted with p-value < 0.001 has been set to 1. The alpha values for the remaining voxels has been set in proportion to r on linear scale ranging from 0 to 1. The lower (conv-1, conv-2, conv-3), intermediate (conv-4, conv-5) and higher (fc-6, fc-7) layer of the ConvNet were found to be optimal for voxels in posterior, intermediate and anterior areas of the visual cortex respectively. This shows that the hierarchy of visual representations in the ConvNet mimics the hierarchy of visual representations in the human brain.

to the low, mid and high-level visual features represented by the brain by a linear transformation. From this it can be concluded that the hierarchy of visual representations in the ConvNet mimics the hierarchy of visual representations in the human brain.

3.3 Investigating Visual Representations in the Human Brain

Insights into visual representations of the human brain can be developed by visualizing the features represented by individual voxels. For this, we used the ConvNet model to predict BOLD activity of individual voxels to a collection of more than 280K natural images (see supplementary material for a detailed description of this image collection). Then for every voxel independently, these images were rank-ordered according to the predicted BOLD activity. The top and bottom images within this ranking provide qualitative intuition about the features that are represented by a particular voxel (see supplementary material for more details). Figure 4 shows the top and bottom six images for two voxels in V1, V4, FFA and PPA. As there were too many voxels to visualize, only two sample voxels from each ROI were chosen in the following way: For each ROI, all voxels predicted with a p-value < 0.0001 were pooled across the four subjects. From this set, two voxels were manually chosen to illustrate the range of visual representations in a ROI. A random sample of voxels from V1, V4, FFA and PPA is shown in the supplementary materials.

The V1 voxels are predicted to increase activity when images consist of high texture and to decrease activity when images contain landscapes. This result is not surprising because V1 is known to contain neurons that respond to oriented edges and images with high texture are likely to excite a large number of V1 neurons. This in turn would cause the V1 voxels to elicit large responses to textured images. The FFA voxels are predicted to increase activity when images contain human and animal faces and to decrease activity when they contain scenes. These results are consistent with previous accounts of FFA [5, 6]. The PPA voxels are predicted to increase activity when images contain scenes and trains and to decrease activity when they contain animate objects. The geometric structure of trains is not very different from that of buildings. This suggests PPA voxels encode specific geometric structures useful for identifying scenes/places and are not likely to be selective for any object categories. This interpretation of features represented in PPA is consistent with the findings of [22] and previous accounts of PPA [8]. These results demonstrate that using the ConvNet model, results of multiple previous fMRI studies that investigated visual representations in individual ROIs [5, 8, 7, 3] can be reproduced using only a single fMRI study.

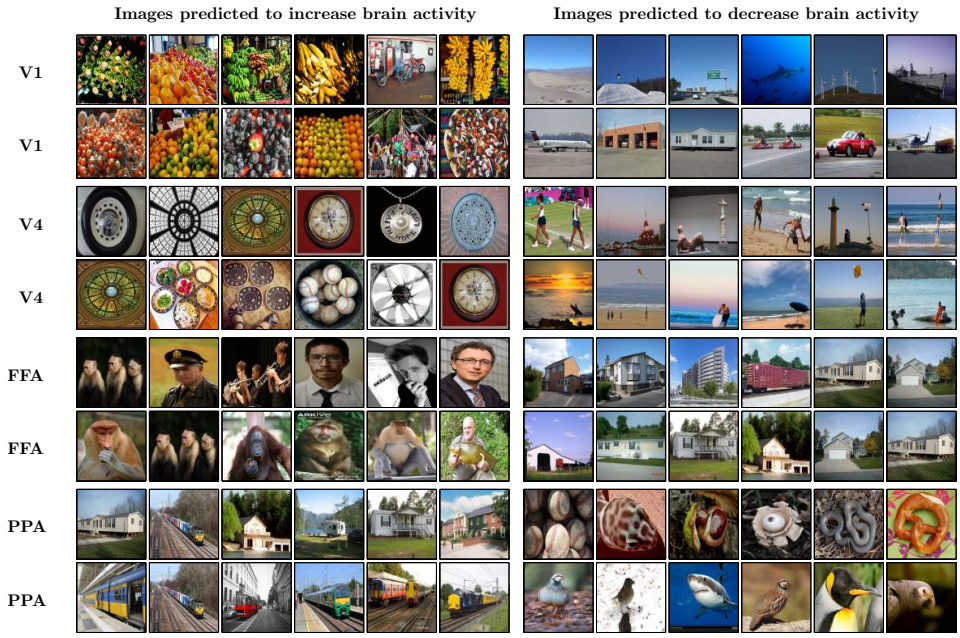

Figure 4: Using the ConvNet model to probe the stimulus tuning of individual voxels. Each row represents the tuning of a single voxel. The visual tuning of two voxels from the visual areas V1, V4, FFA and PPA is shown. These voxels were manually chosen from a set of voxels obtained after pooling voxels from all four subjects. The ConvNet model fit to each voxel was used to filter a set of 280K natural images. The six columns at left show the six images that the ConvNet model for each voxel predicts will most increase BOLD activity, and the six columns at right show the images that the model predicts will most decrease BOLD activity. The V1 voxels are predicted to increase activity when images consist of high texture and to decrease activity when images contain landscapes. The V4 voxels are predicted to increase activity when images contain a circular shapes and to decrease activity when they contain landscapes. The FFA voxels are predicted to increase activity when images contain human and animal faces and to decrease activity when they contain scenes. The PPA voxels are predicted to increase activity when images contain scenes and trains and to decrease activity when they contain animate objects.

Despite several past studies, the understanding of visual representations in V4 is unsatisfactory [9, 23, 24]. Our analysis reveals that a subset of V4 voxels are predicted to increase BOLD activity when images contain a circular shapes and to decrease activity when they contain landscapes. This result is qualitatively consistent with neurophysiological reports that area V4 is selective for curvature and radial patterns [24] and shows that ConvNet can be used to investigate visual representations in intermediate visual ROIs.

4 Discussion

Understanding how the brain transforms low-level visual features into high-level visual features requires developing computational theories that make testable predictions about visual representations in the brain. In the past, such theories have either been based purely on the neurophysiological findings or have been inspired by Barlow's redundancy reduction hypothesis [25]. Hubel and Wiesel's finding of simple and complex cells [2] led to the computational hypothesis that the hierarchy of visual features in the brain was constructed by consequent stages of linear filtering, pooling and point-wise non-linearities. This idea was first championed by the Neocognitron model [26] and later by the HMAX model [27]. In a different line of work, past studies found that computational models based on Barlows idea predicted what features were represented by neurons in V1 [28, 29]. Since then, several studies have attempted to use the redundancy reduction hypothesis for building computational models that explain features represented in visual areas beyond V1 [30, 31]. However, these

studies have met with limited success and no prior study has been able to construct a computational model that provides plausible predictions about features represented across the visual cortex.

In this work we demonstrated that hierarchy of visual representations in the ConvNet mimics the hierarchy of visual representations in the human brain. In contrast to past studies that proposed a similar model of computation [26, 27], the key difference is that the ConvNet model was trained for the task of object recognition. Models used in these previous studies were not optimized for performing any particular task and it is likely that feature representations of these models simply captured natural image statistics. This suggests that computational theories relying only on natural image statistics (i.e. unsupervised learning) maybe insufficient in providing good hypotheses about how the brain represents visual information. As an additional support to this claim we have presented results in the supplementary materials that show that ConvNet outperform a feature descriptor called as Fisher Vectors [32] on the task of predicting brain activity. Fisher vectors capture higher order natural image statistics and were the state of art feature descriptor for computer vision tasks [32] before the advent of deep neural networks in 2012 [17]. While the critics may argue that FV is not the optimal representation for natural image statistics, alternative methods such as autoencoders [33] and boltzmann machines [34] have not been shown to work on complex real world imagery.

These facts taken together suggest that computational models developed for solving the same tasks that the human visual system performs can provide good hypothesis about how the brain processes and represents visual information. In the hindsight, this is not surprising. The representations of a complex information processing system such as the brain must depend on the tasks that it performs. The visual system of the brain is not only adept at object recognition but is also involved in motor tasks such as navigation and manipulation of tools. This suggests that computational models that seek to jointly optimize visual and motor tasks may lead to a better understanding of the human visual system. Some recent works such as [35, 36, 37] that have proposed models for solving visuomotor tasks provide interesting directions for future research in this area.

Some recent studies [38, 39] have provided evidence that ConvNets can explain visual representations in the Inferior Temporal (IT) cortex of macaques and humans. However, these results are not surprising because IT appears to represent semantically meaningful features such as faces and places and the ConvNet was trained for object recognition. What our results show is that the ConvNet mimics the hierarchy of visual representations across the visual cortex. This implies that not only is the ConvNet plausible model of visual processing but it can also be used to study visual representations throughout the visual cortex. Such claims cannot be made based on the results of any previous work.

One potential critique of our work is that unlike the brain, the ConvNet has no feedback or recurrent connections. How is it then that the ConvNet is able to predict brain activity across the visual cortex? One explanation is provided by past studies that have shown that the brain can perform object recognition using feed-forward computations only [40]. Moreover, although the ConvNet model outperforms previously proposed models for predicting brain activity, there still is substantial amount of variance in brain activity that is not explained by the ConvNet model (see table 1).

Another potential critique of our work is that several architectural choices involved in designing the ConvNet (such as the number of layers) were simply made as a result of the fact that they led to good performance on the task of object recognition [17]. These choices may not be optimal for predicting brain activity and consequently the ConvNet model we used is probably sub-optimal. Modifying the ConvNet architecture to incorporate computational mechanisms like recurrence and feedback, and optimally choosing parameters such as the number of layers, the number of units in a layer, and the choice of specific non-linearity will lead to models that make more accurate predictions of brain activity. Future work on developing such models is likely to provide a more nuanced understanding of how the brain processes and represents visual stimuli.

5 Conclusion

The main result of our work is that the hierarchy of visual representations in the ConvNet mimics the hierarchy of visual representations in the human brain. This suggests that understanding visual representations in the ConvNet can help us understand the visual representations in the human brain. As evidence, we have shown that the ConvNet model reveals visual features represented by individual voxels. Our results also provide evidence that computational models optimized for executing

ecologically relevant tasks (like object recognition, performing actions) as opposed to models optimized solely for estimating natural image statistics can provide better hypothesis about how brain transforms low-level visual representations into high-level visual representations.

References

[1] M Spiridon, B Fischl, and N Kanwisher, "Location and spatial profile of category-specific regions in human extrastriate cortex," *Human Brain Mapping*, vol. 27, no. 1, pp. 77–89, 2006.

[2] David H Hubel and Torsten N Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology*, vol. 148, no. 3, pp. 574–591, 1959.

[3] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, no. 7185, pp. 352–355, 2008.

[4] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant, "Reconstructing visual experiences from brain activity evoked by natural movies," *Current Biology*, vol. 21, no. 19, pp. 1641–1646, 2011.

[5] N Kanwisher, J McDermott, and M M Chun, "The fusiform face area: a module in human extrastriate cortex specialized for face perception," *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, vol. 17, no. 11, pp. 4302–4311, 1997.

[6] T Çukur, A G Huth, S Nishimoto, and J L Gallant, "Functional subdomains within human FFA.," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 33, no. 42, pp. 16748–66, Oct. 2013.

[7] I Gauthier, M J Tarr, J Moylan, P Skudlarski, J C Gore, and A W Anderson, "The fusiform "face area" is part of a network that processes faces at the individual level," *Journal of Cognitive Neuroscience*, vol. 12, no. 3, pp. 495–504, 2000.

[8] R Epstein and N Kanwisher, "A cortical representation of the local visual environment," *Nature*, vol. 392, no. 6676, pp. 598–601, 1998.

[9] Anitha Pasupathy and Charles E Connor, "Responses to contour features in macaque area v4," *Journal of Neurophysiology*, vol. 82, no. 5, pp. 2490–2502, 1999.

[10] S V David and J L Gallant, "Predicting neuronal responses during natural vision," *Network*, vol. 16, no. 2-3, pp. 239–260, 2005.

[11] T Naselaris, K N Kay, S Nishimoto, and J L Gallant, "Encoding and decoding in fMRI," *NeuroImage*, vol. 56, no. 2, pp. 400–410, 2011.

[12] D Stansbury, T Naselaris, and J Gallant, "Natural scene statistics account for the representation of scene categories in human visual cortex.," *Neuron*, vol. 79, no. 5, pp. 1025–34, Sept. 2013.

[13] A G Huth, S Nishimoto, A T Vu, and J L Gallant, "A continuous semantic space describes the representation of thousands of object and action categories across the human brain.," *Neuron*, vol. 76, no. 6, pp. 1210–24, Dec. 2012.

[14] Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just, "Predicting human brain activity associated with the meanings of nouns," *science*, vol. 320, no. 5880, pp. 1191–1195, 2008.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.

[16] Y LeCun, B Boser, J S Denker, D Henderson, R E Howard, W Hubbard, and L D Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, 1989.

[17] A Krizhevsky, I Sutskever, and G E Hinton, "Imagenet classification with deep convolutional neural networks.," in *NIPS*, 2012.

[18] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014*, pp. 818–833. Springer, 2014.

[19] Pulkit Agrawal, Ross Girshick, and Jitendra Malik, "Analyzing the performance of multilayer neural networks for object recognition," in *Computer Vision–ECCV 2014*, pp. 329–344. Springer, 2014.

[20] Y Jia, "Caffe: An open source convolutional architecture for fast feature embedding," `http://caffe.berkeleyvision.org/`, 2013.

[21] Richard B. Buxton, *Introduction to functional magnetic resonance imaging book pack: Principles and techniques*, Cambridge University Press, 2002.

[22] Reza Rajimehr, Kathryn J Devaney, Natalia Y Bilenko, Jeremy C Young, and Roger BH Tootell, "The "parahippocampal place area" responds preferentially to high spatial frequencies in humans and monkeys," *PLoS biology*, vol. 9, no. 4, pp. e1000608, 2011.

[23] Anitha Pasupathy and Charles E Connor, "Population coding of shape in area v4," *Nature neuroscience*, vol. 5, no. 12, pp. 1332–1338, 2002.

[24] Jack L Gallant, Charles E Connor, Subrata Rakshit, James W Lewis, and DAVID C Van Essen, "Neural responses to polar, hyperbolic, and cartesian gratings in area v4 of the macaque monkey," *Journal of neurophysiology*, vol. 76, no. 4, pp. 2718–2739, 1996.

[25] Horace B Barlow, "Possible principles underlying the transformations of sensory messages," 1961.

[26] Kunihiko Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[27] Maximilian Riesenhuber and Tomaso Poggio, "Hierarchical models of object recognition in cortex," *Nature neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[28] A J Bell and T J Sejnowski, "The "independent components" of natural scenes are edge filters.," *Vision research*, vol. 37, no. 23, pp. 3327–38, Dec. 1997.

[29] B Olshausen and D J Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, 1996.

[30] Charles F Cadieu and Bruno A Olshausen, "Learning intermediate-level representations of form and motion from natural movies," *Neural computation*, vol. 24, no. 4, pp. 827–866, 2012.

[31] Yan Karklin and Michael S Lewicki, "Learning higher-order structures in natural images," *Network: Computation in Neural Systems*, vol. 14, no. 3, pp. 483–499, 2003.

[32] J Sánchez, F Perronnin, T Mensink, and J Verbeek, "Image Classification with the Fisher Vector: Theory and Practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, June 2013.

[33] Hervé Bourlard and Yves Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological cybernetics*, vol. 59, no. 4-5, pp. 291–294, 1988.

[34] Ruslan Salakhutdinov and Geoffrey E Hinton, "Deep boltzmann machines," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 448–455.

[35] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel, "End-to-end training of deep visuomotor policies," *arXiv preprint arXiv:1504.00702*, 2015.

[36] Pulkit Agrawal, Joao Carreira, and Jitendra Malik, "Learning to see by moving," *arXiv preprint arXiv:1505.01596*, 2015.

[37] Dinesh Jayaraman and Kristen Grauman, "Learning image representations equivariant to ego-motion," *arXiv preprint arXiv:1505.02206*, 2015.

[38] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte, "Deep supervised, but not unsupervised, models may explain it cortical representation," *PLoS computational biology*, vol. 10, no. 11, pp. e1003915, 2014.

[39] Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo, "Deep neural networks rival the representation of primate it cortex for core visual object recognition," *PLoS computational biology*, vol. 10, no. 12, pp. e1003963, 2014.

[40] Simon Thorpe, Denis Fize, Catherine Marlot, et al., "Speed of processing in the human visual system," *nature*, vol. 381, no. 6582, pp. 520–522, 1996.