Characterizing the Visual properties of intermediate layers of Deep Vision models that are predictive of V4

Hari Bandi
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139
hbandi@mit.edu

Jambay Kinley
Harvard College
Harvard University
Cambridge, MA 02138
j_kinley@college.harvard.edu

Abstract

Recent studies have shown that deep vision (neural network) models trained for tasks such as the ImageNet challenge are highly predictive of neural activity in cortical areas in the visual stream and behavior. Although lots of research has been done on studying the low level regions like V1 and higher-level regions like IT, less work has been done to understand the intermediate layers like the V4. Based on previous works on the V4 and methods used to characterize the response properties of unit(s) in neural networks by generating stimuli that maximally stimulate them, we will attempt to characterize the visual properties of layers that are highly predictive of V4 neural activity in multiple computer vision models. We expect to find features that are not as high level as those in the later layers and similar to those previous studies on the V4 have suggested like color areas and curvatures. The similarities and differences in the response properties of different models might reveal to us some insight into the computations happening in the intermediate layers in the ventral visual stream. The results could also be used to guide future studies in the actual visual stream by allowing us to test different candidate models by creating stimuli that elicit different responses from the models.

1 Introduction

The ventral visual system has been known to be made of multiple distinct areas starting from the V1 to the IT with largely feed forward connections supplemented by feedback connections. Although, lots of research has been done on studying the low level regions like V1 and higher-level regions like IT, relatively less work has been done to understand the intermediate layers like the V4.

Early visual areas, such as the V1 cortex, capture low-level features such as edges and center-surround patterns [1; 2]. In contrast, the highest ventral visual areas of the inferior temporal (IT) cortex, can be used to decode object category and are robust to significant variations present in natural images [3; 4]. The visual properties of mid-level visual areas such as V2 and V4 are less well understood, but these areas appear to contain intermediate computations between simple edges and complex objects. Although it is also crucial for visual attention, our focus in this study will be on the its role in feature processing. Roe et al. [5] note that a diversity of responses including selectivity for color and orientation have been found in V4. The V4 in monkeys have been shown to be not homogeneous and that it may consist a collection of modules [6; 7].

Although the functional complexity of the V4 poses a hard challenge to physiological studies of the area, recent advances in computer vision models [8; 9; 10] and methods for characterizing [11; 12] and comparing the response properties of these models with the human brain provide us with an promising avenue for gaining more insight into the visual system. Recent studies have shown that deep vision (neural network) models trained for tasks such as the ImageNet challenge [13] are highly predictive of cortical areas in the visual stream and behavior [14].

Schrimpf et al. [15] developed the Brain-Score metric which benchmarks how closely an Artificial neural network for object recognition resembles the Brain by measuring the model's predictivity of neural and behavioral data. All of the models the authors tested achieved scores above .6 out of a possible score of 1 (the max achievable is .892). The best performing model was found to be VGG-19 [8] (layer *block3_pool*) with a score of .672.

Preliminary results released in a recent memo from the Center for Brain Mind and Machines show that there are many one-to-one mappings between single units in a deep neural network model and neurons in the brain [16]. These mappings have been found in many state-of-the-art deep NNs with the best mappings found in models with high performance on the Imagenet challenge. The highest-scoring model overall was found to be NASNet-Large [9] with a correlation of 0.6224 for V4, and 0.4337 for IT.

Inspired by one of the questions posed by Yamin et al [17] where the authors note the possibility of performing "high-throughput virtual electrophysiology to characterize the [computational] model's internal structure" to gain insight into tuning curves of the corresponding cortical areas, in this paper we attempt to characterize the visual properties of the intermediate layers of various computer vision models using max image visualization techniques and analyzing the responses of the layers to these stimuli.

Owing to space and time constraints we will focus on a few filters in the *Cell_0* layer of NASNet-large. Individual units in two of these filters have been found to be highly correlated with specific sites in the V4. We compare these filters with other filters which haven't been found to be correlated (at least for now) to see if there are any interesting differences. We will also show some of the filters found in the *block3_pool* layer of VGG-19 which received the highest score on the Brain-Score metric.

We first describe the methods used to generate the max images and present some of the stimuli thus found. We will then analyze the max image and response of filter 804 in the *Cell_0* layer of NASNet-large which has high correlation with a V4 site (r = 0.6463). We will then compare this filter with filter 783 which is also correlated with another V4 site (r = 0.3515) and filters 444 and 567 which haven't been matched so far.

2 Methods

2.1 Generating Maximal Images

Understanding the visual properties captured by deep Convolutional Models has become an important field of research due to the growing popularity and success of such models. As such, lots of work has already been done and the primary method used in this paper is based on them.

In previous work, Erhan et al. [11] synthesized images that maximally activate neuron(s) of interest in pretrained convolutional networks. In their method, they first start with some initial input image $\mathbf{x} = \mathbf{x}_0$. Consequently, the activity $a_i(\mathbf{x})$ of the neuron/unit i of interest is computed and the image $\mathbf{x}$ is altered by taking steps in the direction of the gradient $\partial a_i(\mathbf{x})/\partial \mathbf{x}$ until some acceptable termination condition. This gradient-based process generates inputs that activate the specific neuron/unit with each iteration of the gradient method. Althought this method achieves to generate inputs that maximally activate a neuron, it often tends to generate non-natural (which do not resemble natural images) looking images as described by Yosinki et al. [18] In order to overcome some of these shortcomings, we complemented the gradient-based methods of generating maximal images by adding two of the regularization methods suggested by Yosinki et al. The regularization is achieved via a regularization operator $r_\theta(\cdot)$ that maps an input $\mathbf{x}$ to its regularized version. Therefore, the gradient-based methods with regularization simplify to the following:

$$\mathbf{x}^{(t+1)} = r_\theta \left(\mathbf{x}^{(t)} + \eta \partial a_i(x)/\partial \mathbf{x} \right)$$

In this work, we use the following regularization operators proposed in Yosinki et al. to synthesize maximal images for neurons/units of interest,

- The first method of regularization used is L_2 decay where large values in $\mathbf{x}$ are penalized by an operator

  $$r_\theta(\mathbf{x}) = (1 - \theta_{\texttt{decay}})\mathbf{x}$$

  where $\theta_{\texttt{decay}}$ is a hyperparameter that lies between 0 and 0.5.

- The second method is using Gaussian blur to penalize high frequency information. This is achieved by convolving the image with a blur kernel.

$$r_{blur}(\mathbf{x}) = \texttt{GaussianBlur}(\mathbf{x}, \theta_{\texttt{b\_width}})$$

  where $\theta_{\texttt{b\_width}}$ is the standard deviation of the Gaussian kernel applied.

Yosinki et al. recommend three other types of regularization but we do not use them since they found the first two methods to be enough for most uses.

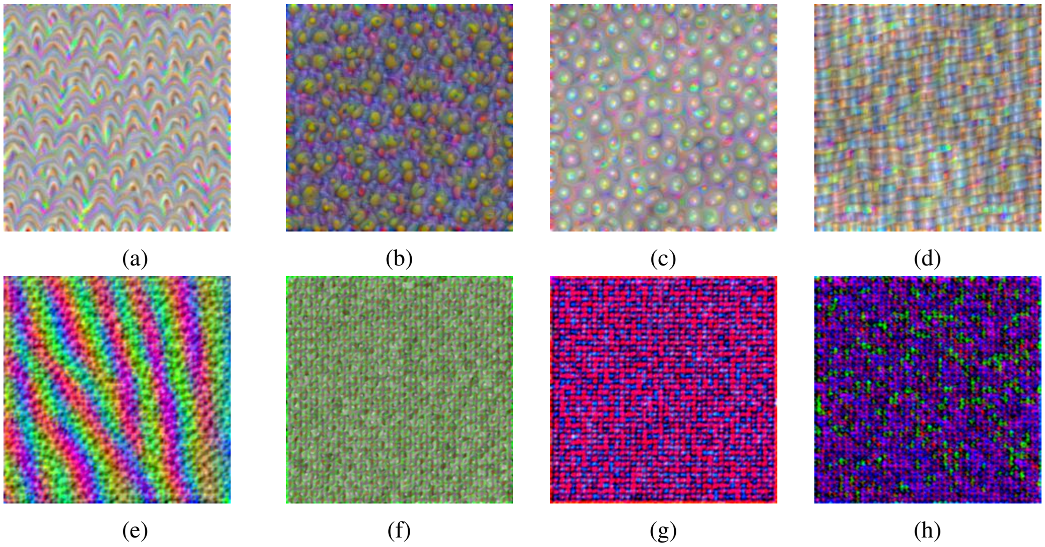
Although our initial plan was to get maximal images for individual units in the layer of interest, we found that finding maximal images for each layer by maximizing the average activity of the layer produced similar images and was much easier to synthesize due to more gradient information. The maximal image of a layer is a tessellated version of the smaller, maximal images of the units so the results do not change.

Our code base can be found **here**. It is written in python using the PyTorch and Numpy libraries. Most of the code is modified from a **CNN Visualization repository** by Utku Ozbulak with the gaussian blur functions borrowed from the **Deep Visualization Toolbox** built by Yosinki et al.

We use a pytorch port of NASNet-Large (with original weights) made available by Cadene on his **github repo**. and a pytorch port of the VGG-19 model available as a part of the **torchvision library**.

3 Results

We illustrate some examples of the maximal images we synthesized in Figure 1. As can be seen, the images generated are diverse but the structure is regular within each image.


(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 1: Maximal images for different filters a) 7, b) 57, c) 13 and d) 165 in the *block3_pool* layer of VGG-19 and filters e) 444, f) 567, g) 783 and h) 804 in the *cell_0* layer of NASNet-Large.

Unit $(804, 21, 19)$ in the *Cell_0* layer of NASNet large has been found to be highly correlated with a V4 site in [16]. We intend to analyze and characterize most of the filters and layers but we used filter 804 as a sample for analysis in this paper.

3.1 Analysis of Filter *Cell_0, 804*

The stimuli in [16] were grayscale images so we do not expect to get much, if any, information about color selectivity from the results. However, to allow for tests of color selectivity of the filters, we also synthesized images with only one color - red, green, or blue, keeping the other two channels 0.

Figure 2 shows the maximal images with all colors and singular colors. A qualitative visual analysis of the maximal images suggest that the maximal stimuli for *Cell_0, 804* is made of grid like patterns.

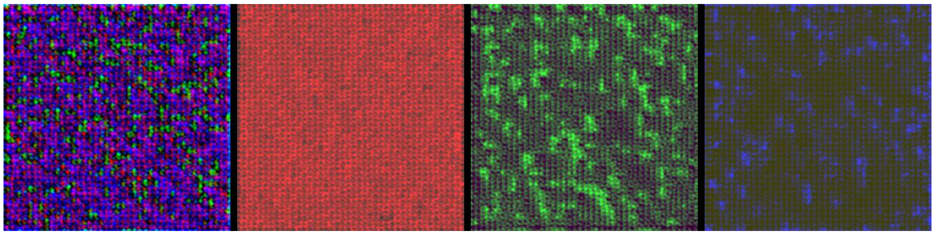


Figure 2: Maximal Images with all color channels and individual color channels

In order to get a quantitative overview of the selectivity of this filter for such patterns, we compared the maximum and average activities of the units to different images in Figures 3.

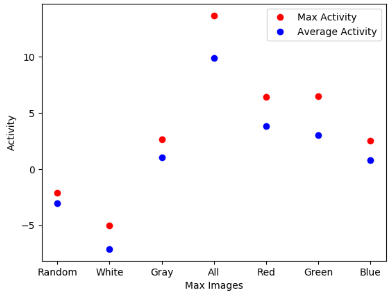
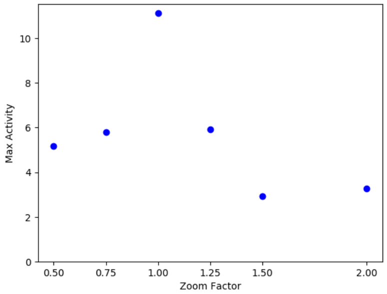


Figure 3: Maximum and Average Activities of the units in layer *Cell_0, 804* to the maximal images. *Random* is a randomly generated image with gaussian noise, *White* is an all-white image and *Gray* is a grayscale version of the first image in Figure 2. The rest are the same images in 2 in their corresponding order

3.1.1 Investigating Selectivity for the specific grid-like pattern

We measured the maximal activation of the layer to the maximal image at different zoom factors keeping the overall image size constant. The results are seen in Figure 4. From this, we observe that the layer responds best when the object of interest is of a certain size. It doesn't fire indiscriminately to all sized objects.



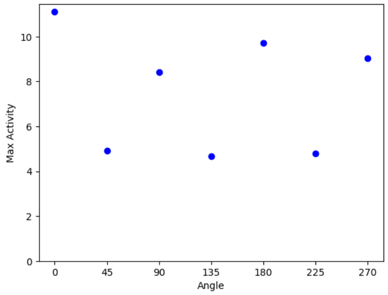


Figure 4: Maximum Activity of layer *Cell_0, 804* to stimuli at different zoom factors

Figure 5: Maximum Activity of layer *Cell_0, 804* to stimuli at different rotation angles

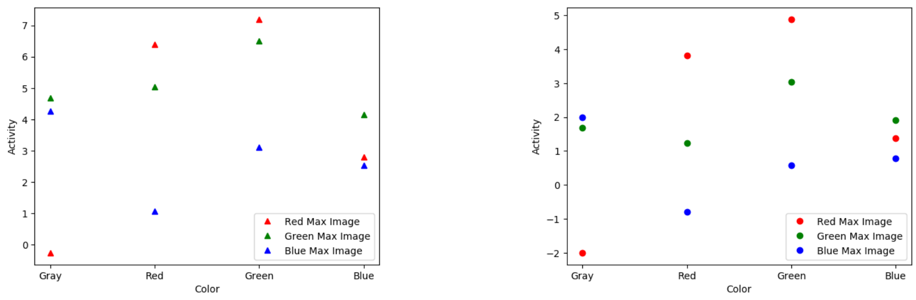
We also observe selectivity to the orientation of the grid. Figure 5 shows that the activation is maximal when the grid aligns with the vertical and horizontal axis of the image.

3.1.2 Investigating Color Selectivity

In Figure 6, we see that the layer's activation was significantly greater for the colored image was significantly higher than those for the grayscale and single colored images. The stimuli for the experiments in [16] were grayscale but the discrepancy in the activities and maximal images for different colors provides us with an interesting question about color selectivity of the layer and it's corresponding V4 site.

To acertain that the discrepancy was not due to the maximal image for the color channels being different from one another, we generated three new images for each color by switching the color on the single colored maximal images to the other two colors as well generating a grayscale version of the maximal image.

The maximum and average activities for each image are plotted in Figure 6. We do not observe significant difference between the red and green channels but the activity is lower for the blue channel. It is significantly smaller for the grayscale version of the red max image. This image has a regular grid structure of uniform color. The results suggest that contrasting colors might be required along with grid-like patterns form maximal actviation.



(a)

(b)

Figure 6: a) Maximum and b) Average activities of layer *Cell_0, 804* to different colored images. “Red Max Image" indicates that the images were generated from the maximal for the color red by switching the active color channel to the one of interest

3.2 Comparison with other layers

Due to space constraints we are unuable to provide equally detailed analysis of other units/layers in the NASNet-Large model but we note the following points. Images for the layers below can be found on our code repository for reference.

3.2.1 Layer *Cell_0, 783*

This layer displays a similar selectivity to grid-like patterns of specific size and orientation. The responses are significant smaller for the green channel just as suggested by the maximal image in 1 h.

3.2.2 Layer *Cell_0, 567*

This layer displays a selectivity to circular green objects arrange in a grid. It is less sensitive to orientation change and highly selective for green color.

3.2.3 Layer *Cell_0, 444*

This layer has a very different maximal image compared to the layers we have seen so far. As seen in Figure 1 e. it is made of thick colored stripes. The color or orientation of the stripes do not significant affect the activity. Like *Cell_0, 783* does not prefer low contrast gray-scale images.

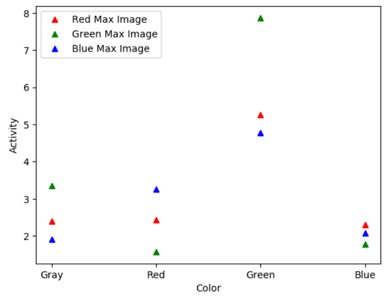


Figure 7: Maximum activity of layer *Cell_0, 567* to different colored images.

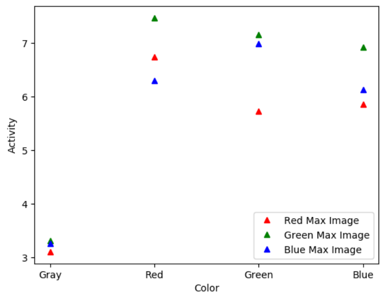


Figure 8: Maximum activity of layer *Cell_0, 444* to different colored images.

3.3 Implications

The results we have shown here are a small part of what is possible through the method of synthesizing maximal images for deep convolutional networks and using them to characterize the visual properties of the model.

The one-to-one mappings recently found between units in deep NNs and sites in the visual cortex could provide more uses for our methods. We can analyze the units found to be highly correlated with individual sites to gain insight into the visual cortex and perhaps guide future studies of the human visual system, especially the V4.

We could compare the visual properties of layers with known mappings with those of the layers currently unmatched. Differences found between them might tell us about the fundamental functional parts of the visual object recognition system. The insights gained through the comparative analysis and the stimuli we can generate could allow us to find new mappings efficiently.

For instance, the results in this paper suggest that some of the layers are highly selective for specific patterns of specific colors. We could find more fine grained mappings by using colored stimuli in experiments similar to those done in memo. Moreover, since a specific unit corresponds responds to a section of the stimuli (image), we might be able to find mappings for V4 sites that have the same response properties but with their receptive fields located at different sections of the stimuli. By probing for these sites, we could learn about the spacial organization of such sites.

4 Future Work

This project has synthesized maximal images for all the filters in a layer of the VGG-19 network and characterized the response of the filters. This is a small but important step in a big project that would require more work on the following:

Analyze and characterize other layers. The first thing to do would be to characterize layers with maximal images similar to those done in this paper. We can set up a set of procedures to systematically and thoroughly study these neurons.

Characterize other models and compare them. As noted in the introduction, one of the most useful applications of our method would be in comparing different computer vision models and use their similarities and differences to gain better insights into the V4. Properties common in all models might suggest important features needed for visual processing while the differences would allow us to test different candidate models for the V4.

5 Conclusion

Area V4 is one of the least studied regions in the visual system mostly due to the complexity of its organization and tuning curves. "Virtual Electrophysiological" studies of deep computer vision models that are highly predictive of the neural activity of the V4 provide a promising avenue for research.

In this project, we used a gradient-based optimization method to generate images that maximally stimulate particular layers of interest and performed an analysis of the response properties of one of the layers. Our results suggest that units in the intermediate layers of the model are selective for colored shapes, size and possible curvatures.

Acknowledgments

We would like to thank the teaching staff of *Science of Intelligence* for supporting our project and organizing the course.

References

[1] Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., ... & Rust, N. C. (2005). Do we know what the early visual system does?. *Journal of Neuroscience*, 25(46), 10577-10597

[2] Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978). Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *The Journal of physiology*, 283(1), 53-77.

[3] Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., & Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature neuroscience*, 11(11), 1352.

[4] Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749), 863-866.

[5] Roe, A.W., Chelazzi, L., Connor, C.E., Conway, B.R., Fujita, I., Gallant, J.L., Lu, H. & Vanduffel, W. (2012) Toward a Unified Theory of Visual Area V4. *Neuron Review*, Volume 74, Issue 1, p12-2. 10.1016/j.neuron.2012.03.011

[6] Conway, B.R., Moeller, S. & Tsao, D.Y. (2007) Specialized color modules in macaque extrastriate cortex. *Neuron*, 56:560–573.

[7] Harada, T., Goda, N., Ogawa, T., Ito, M., Toyoda, H., Sadato, N. & Komatsu, H. (2009) Distribution of colour-selective activity in the monkey inferior temporal cortex revealed by functional magnetic resonance imaging. *Eur J Neurosci*, 30:1960–1970

[8] Simonyan, K. & Zisserman, A. (2015) Very Deep Convolutional Networks for Large-scale Image Recognition. *ICLR*

[9] Zoph, B., Vasudevan, V., Shlens, J. & Le, Q.V. (2017) Learning Transferable Architectures for Scalable Image Recognition. arXiv:1707.07012

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

[11] Erhan, D., Bengio, Y., Courville, A, & Vincent, P. (2009) Visualizing higher-layer features of a deep network. Technical report, Technical report, University of Montreal.

[12] Zeiler, M.D. & and Fergus, R. (2013) Visualizing and understanding convolutional neural networks. arXiv:1311.2901.

[13] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. & Fei-Fei, L. Imagenet: A large-scale hierarchical image database. (2009) *Computer Vision and Pattern Recognition*, IEEE Conference on, pp. 248–255.

[14] Cadieu, C.F., Hong, H., Yamins,D.L.K., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J. & DiCarlo, J,J. (2014) Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*

[15] Schrimpf, M., Kubilius, J., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D.L.K. & DiCarlo, J.J. (2018) Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*, 10.1101/407007

[16] Arend, L., Han, Y., Schrimpf, M., Bashivan, P., Kar, K., Poggio, T., DiCarlo, J.J. & Boix, X. (2018) Single units in a deep neural network functionally correspond with neurons in the brain: preliminary results. *CBMM Memo No. 093*

[17] Yamins, D.L.K. & DiCarlo, J,J. (2016) Eight open questions in the computational modeling of higher sensory cortex. *Current Opinion in Neurobiology*, 37:114–120.

[18] Yosinski, J., Clune, J., Nguyen, A., Fuchs, T. & Lipson, H. (2015) Understanding Neural Networks Through Deep Visualization. *Deep Learning Workshop, International Conference on Machine Learning (ICML).*