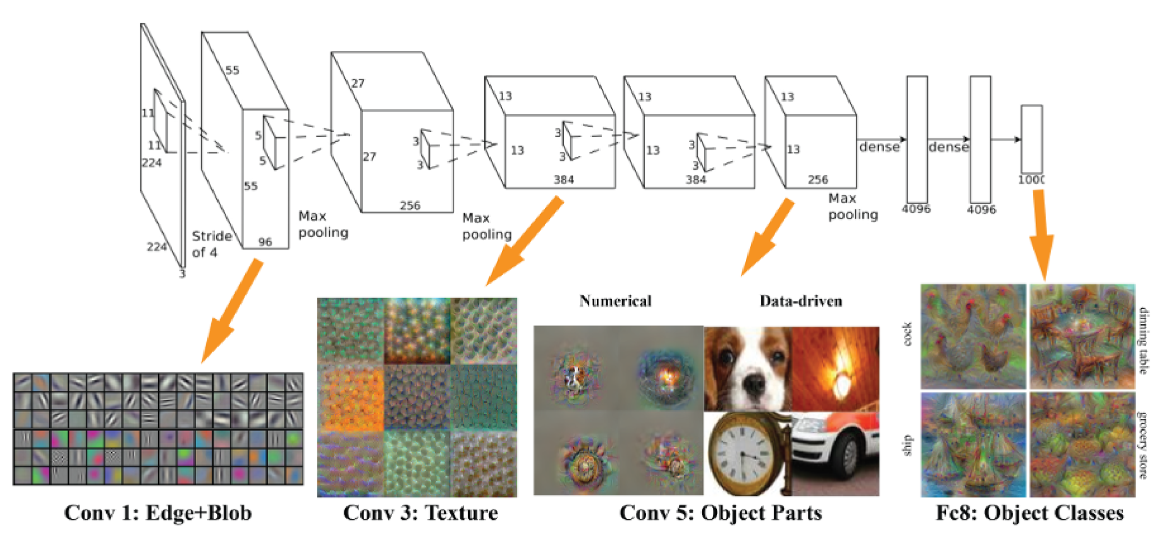
Texture Selectivity in Deep Image Recognition Networks

David Welin Grossman, Md Nasir Uddin Laskar, Luis G Sanchez Giraldo, Odelia Schwartz
Department of Computer Science, University of Miami, FL.
{david.grossman, nasir, lgsanchez, odelia}@cs.miami.edu

Motivation and Contributions

Over the past 5 years, major advances in Machine Learning sparked progress in using convolutional neural networks (CNNs) for image recognition tasks [6]. Although CNNs only loosely mimic the brain hierarchy, recent work showed that CNNs trained on image recognition can predict some properties of visual cortex [5]. For example, in some deep networks, layers in CNNs learn similar representations to areas V1 and the inferior temporal cortex in the brain. Recent work at the University of Miami found that in a limited number of networks, middle layers of deep networks learned texture selectivity that was on par with the brain's texture selectivity in area V2.

CNNs are also used to generate pastiches, artistic works that imitate the style of another [2]. For example, one can take a photograph and make it look like a Monét painting. This author will explore the ability of CNNs to recognize images with textures other than their natural ones.



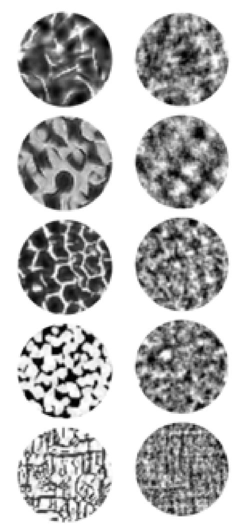
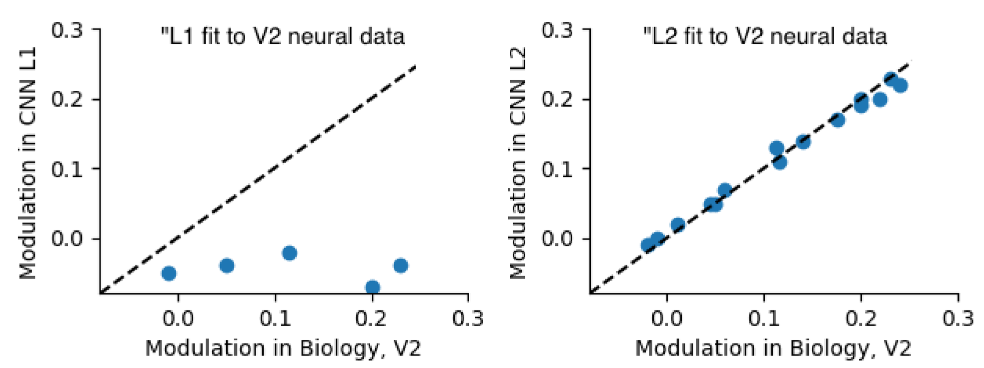


Fig.: *Left*: AlexNet CNN layers and visualization of learned representation (image from MIT CSAIL). *Center*: Example of pastiche generation. Top left image is the original and the other three are Monét-styled pastiches. [1]. *Right*: Texture patches used in [4]. Left column: Natural textures. Right column: Noise textures.

Texture Selectivity in AlexNet-Based Networks

Six networks based on AlexNet were used to analyze the robustness of results from Schwartz, et. al. Natural and noise textures were generated and results were compared by taking the modulation index. **Modulation Index** is computed by taking the difference of the response in naturalistic textures to the response in noise and dividing by their sum. High modulation index indicates a better representation of naturalistic textures. The modified networks were created by removing 1 to 4 layers from AlexNet (8 layers in total). All networks but the one without 4 layers performed well at object recognition tasks (>50%) and at texture selectivity in comparison to biological data. However, the network with 4 layers removed performed relatively well (<30%) but performed poorly at texture selectivity.

AlexNet Original



AlexNet Minus 4 Layers

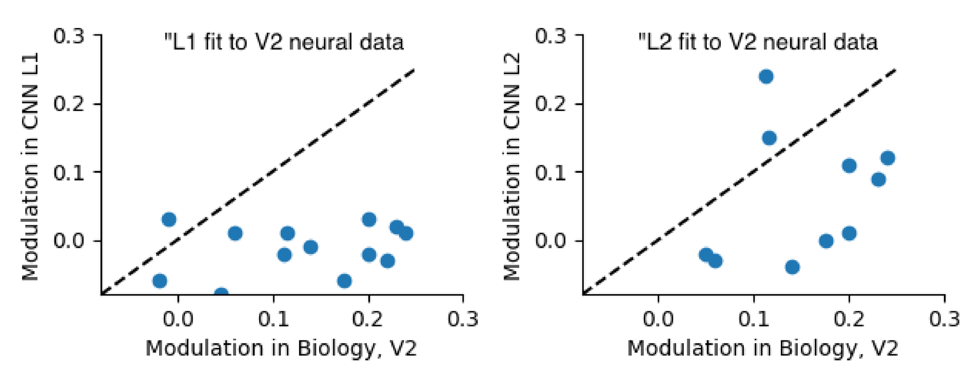
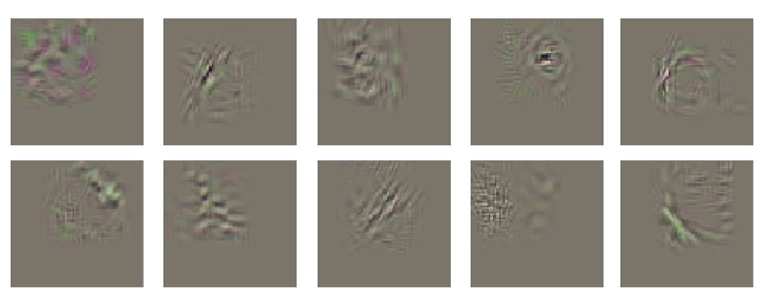


Fig.: The modulation index of Layer 2 (but not Layer 1) units of a CNN are well matched to the V2 neural population data in Macaque brain [3] for a set of 15 natural-noise texture pairs. We find that this breaks down in the reduced AlexNet.

Visualization [7] of the neurons in the original AlexNet compared to the AlexNet without 4 layers varies greatly. Out of 256 neurons, 2 in 10 neurons appears to learns a useful representation in the reduced AlexNet, while most (if not all) neurons in the original Alexnet seem to be useful filters.

AlexNet Original



AlexNet Minus 4 Layers

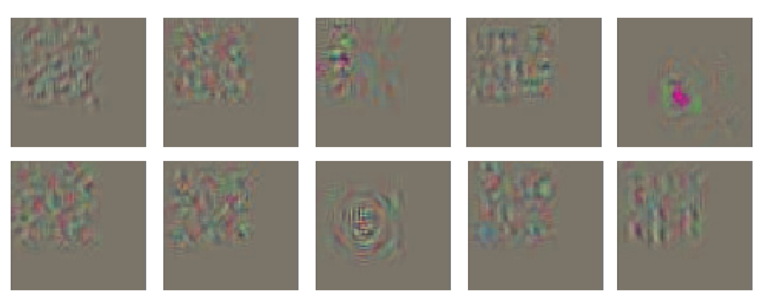


Fig.: Visualization of L2 layers in the original and modified AlexNet.

Object Recognition in Texturized Images

Pastiche images often have the same structure, but have different textures and color schemes. Humans can easily depict the objects in a painting, as long as the painting style isn't too abstract. However, we find that multiple state-of-the-art object recognition systems have difficulties in recognizing pastiche images. Object detectors that attempt to replicate human vision will need to classify images correctly no matter their representation, from line drawing to painting to natural image.

The best image from each class of ImageNet [8] were texturized using a style transfer algorithm [9]. The texturized images were then tested on AlexNet [6], ResNet [10], and VGG-19 [11]. Each deep network classified 150-200 texturized images perfectly based on the distance metric; however the Top-5 accuracies for each network on the texturized images are much lower than their normal recognition accuracy. The **distance metric** refers to the square of the Euclidean Distance between the natural and texturized probability functions.

Accuracy	AlexNet	ResNet-50	VGG-19
Top-1 True	.593	.772	.745
Top-5 True	.818	.933	.920
Top-1 Natural	0.999	0.983	0.978
Top-1 Texturized	0.598	0.569	0.489
Top-5 Natural	1	0.999	0.999
Top-5 Texturized	0.773	0.776	0.701

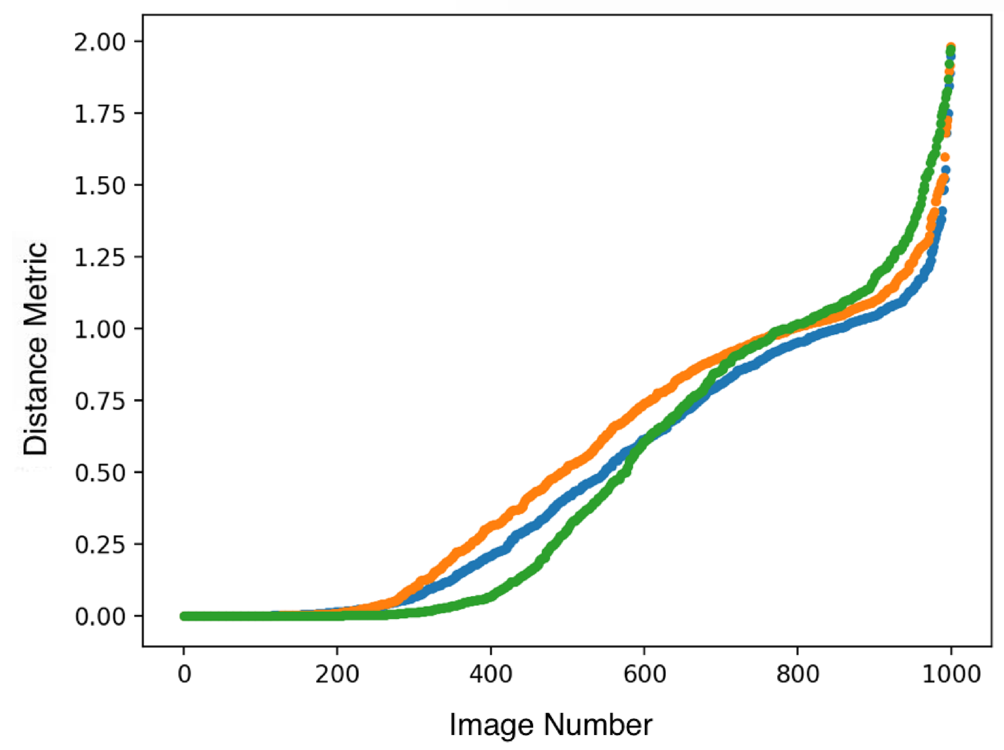


Fig.: *Left* The accuracy of each network when tested on a large dataset (over 20,000 natural images), on a set of 1,000 natural image, and on the same set of 1,000 images after texturization. *Right* The distance metric between the natural and texturized images. Blue: AlexNet, Green: ResNet-50, Orange: VGG-19.

Conclusions and Future Works

- L2 units in AlexNet-based CNNs responded similarly to biological V2 data. However, although the AlexNet-based network with 4 layers removed dropped only partially in object recognition, it did not mimic the biological data in terms of texture selectivity, suggesting the importance of training in a deep network. We are exploring importance of other parameters and specific computations.
- We will explore AlexNet architectures that do not respond to texture and probe their units to see what they do respond to.
- Object Recognition of texturized images is relatively poor when compared to natural images. Future work will focus on understanding why these networks fail to classify texturized images correctly.
- We will focus on building networks that are texture invariant, meaning they will respond similarly no matter the medium of an image. This will begin by training AlexNet on both texturized and natural images.
- We will explore the importance of the ratio between L1's and L2's receptive fields in texture selectivity with Ariel Lavi. We hypothesize that the receptive field must roughly double from the 1st layer before texture selectivity occurs, which is similar to the receptive field ratios between V2 and V1.

References

[1] V. Dumoulin, J. Shlens, M. Kudlur: *ICLR*, 2017.
[2] L. A. Gatys, A. S. Ecker, M. Bethge: *arXiv*, 2015.
[3] C. M. Ziemba, J. Freeman, J. A. Movshon, and E. P. Simoncelli: *PNAS*, 2016.
[4] J. Freeman, C. M. Ziemba, D. J. Heeger, E. P. Simoncelli, and J. A. Movshon: *Nature Neuroscience*, 2013.
[5] N. Kriegeskorte *Cold Spring Harbor Labs*, 2015.
[6] A. Krizhevsky, I. Sutskever, G. E. Hinton: *NIPS*, 2012.
[7] M. Zeiler, and R. Fergus: *ECCV*, 2014
[8] D. J. Dong, et. al.: *Proc. CVPR*, 2009
[9] C. Y. Smith: *GitHub: github.com/cysmith/neural-style-tf*
[10] K. He, X. Zhang, S. Ren, J. Sun: *arXiv*, 2015
[11] K. Simonyan, A. Zisserman: *arXiv*, 2015

This work was supported by a National Science Foundation REU grant and with support from the University of Miami Computer Science Department and Center for Computational Sciences.