Texture Selectivity in Deep Convolutional Neural Networks

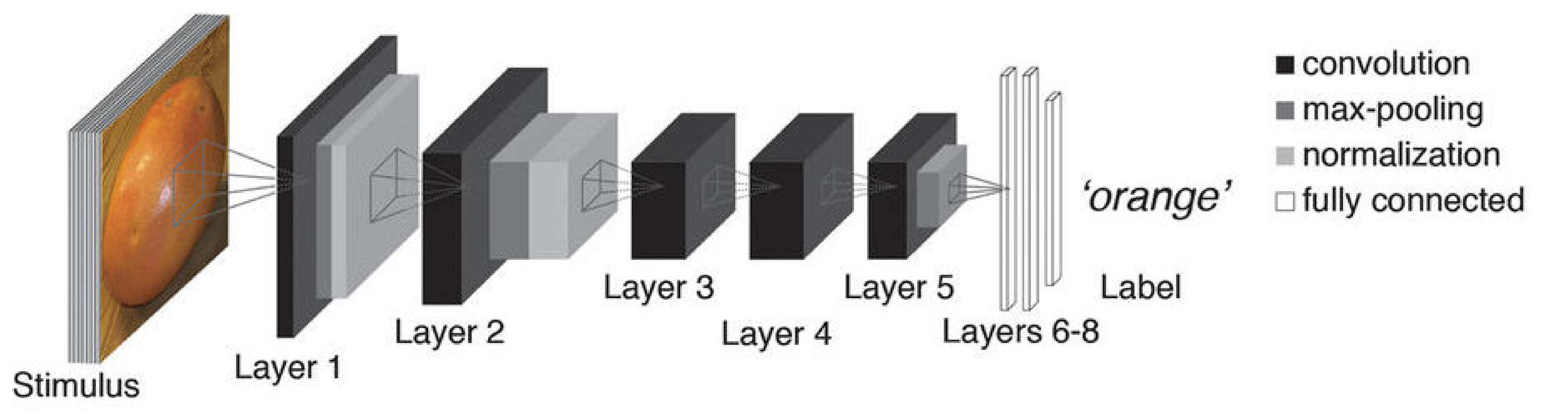
Ariel Lavi, Md Nasir Uddin Laskar, Luis G Sanchez Giraldo, Odelia Schwartz
Department of Computer Science, University of Miami
ariellavi@ucla.edu, {nasir, lgsanchez, odelia}@cs.miami.edu

UNIVERSITY OF MIAMI
DEPARTMENT of COMPUTER SCIENCE

Background and Motivation

In the past five years, there has been significant progress in using convolutional neural networks (CNNs) for image recognition tasks. This progress was initiated by a 2012 publication which featured an eight-layer CNN architecture known as AlexNet [4]. CNNs are loosely inspired by the structure and hierarchy of the visual pathway in the brain. Interestingly, recent research has shown that deep CNNs trained for image recognition can predict certain response properties of visual cortical neurons [8] [3].

Biological studies on macaques suggest secondary visual cortical (V2) selectivity to textures that was not found in the primary visual cortex (V1) [2] [1]. Recent work at University of Miami by Schwartz et al. demonstrated that layer 2 (L2) units in AlexNet develop texture selectivity that provides an excellent fit to the macaque V2 data. To demonstrate the robustness of these results, the author used similar analysis techniques on variations of AlexNet.



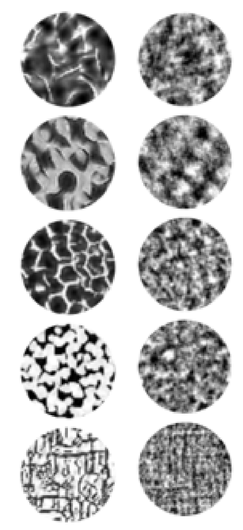


Fig.: *Left*: AlexNet CNN layers (image from MIT CSAIL). *Right*: Sample patches from six of the fifteen texture categories used in [2]. Left column: Natural textures. Right column: Noise textures.

Texture Selectivity in Variations of AlexNet

Four variations of AlexNet were created and trained, and the L2 texture selectivity was analyzed using the same methods from Schwartz et al. The variations were created by changing hyperparameters of the first convolutional and pooling layers. Specifically, the hyperparameters were tuned in order to give a different ratio of L2 to L1 receptive field size (hereafter referred to as R). In biology, this ratio is approximately equal to 2. For Variation 4, the first normalization layer was removed in order to observe the effects of normalization on texture selectivity. The normalization layer in CNNs was designed to mimic biological local response normalization in the primary visual cortex (V1). Though normalization is thought to be important in V1, its effect on higher visual areas is not well understood.

Fifteen naturalistic texture images and fifteen spectrally matched noise images were generated. To quantify results, the modulation index was calculated for each variation of AlexNet. The **Modulation Index** is computed by taking the difference of the responses to naturalistic textures and noise textures, and dividing by the sum of the responses. A higher modulation index indicates a larger differential response from textures to noise, and implies increased selectivity to texture.

AlexNet Original | **AlexNet Variation 1** | **AlexNet Variation 4**

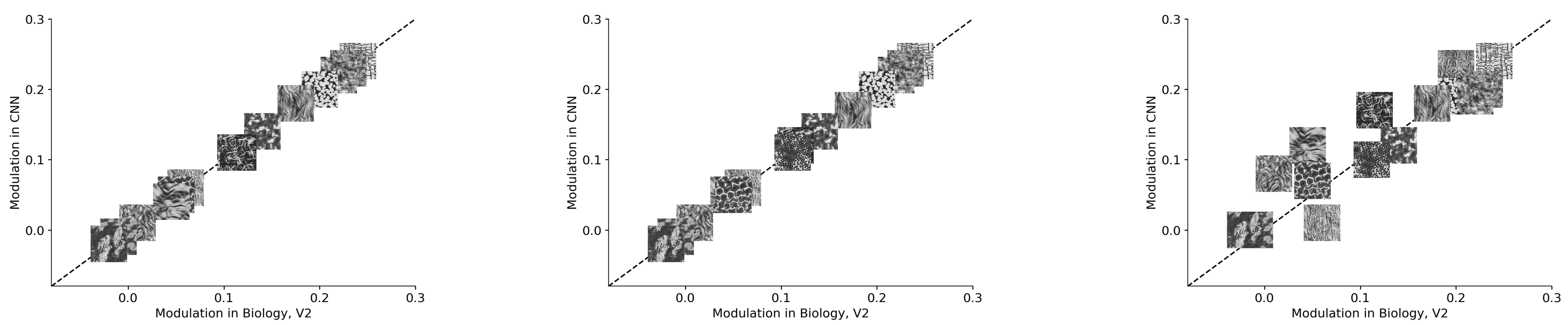


Fig.: The modulation index of Layer 2 units of AlexNet are well matched to the V2 neural population data in Macaque brain [1] for a set of 15 natural-noise texture pairs. This appears to break down upon removal of the first normalization layer (variation 4). *Left*: Original AlexNet, $R \approx 3.53$. *Middle*: AlexNet Variation 1, $R \approx 2.04$. *Right*: AlexNet Variation 4, $R \approx 2.57$.

For image recognition accuracy, the four AlexNet variations performed similarly to the original AlexNet. All four variations had a top-1 accuracy higher than 50%, whereas the original AlexNet had a top-1 accuracy of 57.1% [4]. Interestingly, Variation 4 (normalization removed) indeed shows a reduced fit to biological V2 data. The original AlexNet was trained on 360,000 iterations of the ImageNet dataset[6], whereas the four variations were trained on 220,000 to 270,000 iterations.

Visualization with Deconvolution

With the rise in use of deep CNNs for image recognition, there grew a need to better understand how they achieved such improved accuracy. Zeiler and Fergus [5] developed a visualization technique that gives insight into the workings of the middle layers of CNNs by revealing the input stimuli that excite feature maps for a given layer. The technique makes use of Deconvolutional Neural Networks, which use either the inverse or transpose of the same operations found in CNNs. Whereas CNNs map pixelated inputs (images) into feature maps, Deconvolutional Neural Networks map layer feature maps back into pixel space.

Here, the author uses visualization with deconvolution to give qualitative insight into the texture selectivity of L2 for the different trained variations of AlexNet. The texture and noise images used in the previous section were forwarded through each CNN. The visualizations below show the responses of L2 units.

Sample | **AlexNet Original** | **AlexNet Variation 1** | **AlexNet Variation 4**

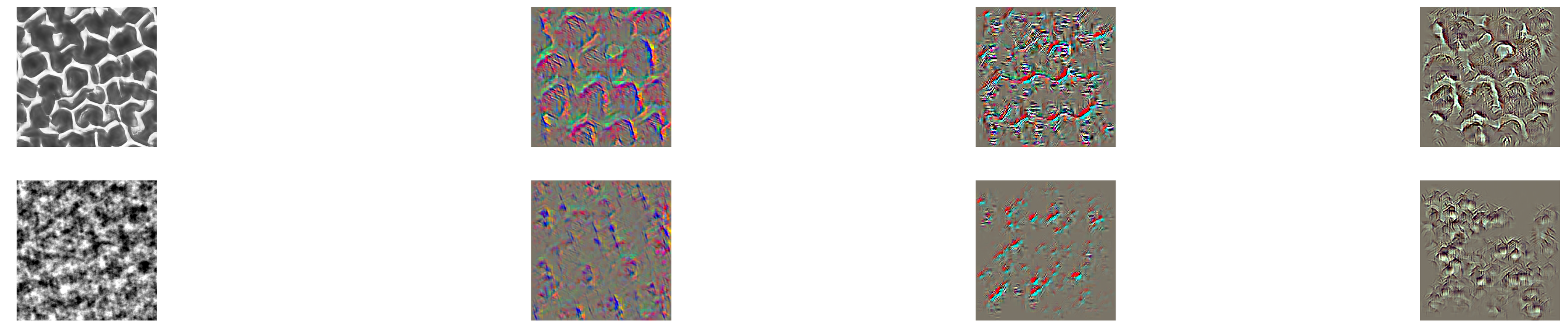


Fig.: *Top*: Naturalistic texture image sample, and visualizations of associated L2 activation. *Bottom*: Spectral noise image sample, and visualizations of associated L2 activation.

Conclusions and Future Works

- Variations of AlexNet displayed image recognition accuracy similar to the original AlexNet (>50%). The responses of L2 units of AlexNet variations to texture and noise images provided a close fit to the biological V2 data, with the exception of variation 4 (normalization removed). This suggests that normalization may have an impact on texture selectivity in L2 and higher.
- There was no evidence of a correlation between the ratio R and image recognition accuracy in the variations of AlexNet.
- Future assignments should continue to explore variations of AlexNet with different R values in order to generate a more robust dataset. On the other hand, multiple variations of AlexNet can be generated for the same R by tuning hyperparameters of the L2, in order to decouple the effects of changing R from effects of changing other hyperparameters.
- Future experiments can involve the use of different normalization methods in AlexNet which can be tested alongside a variation of AlexNet with normalization removed.

Acknowledgements

Many thanks to Dr. Odelia Schwartz for her support, guidance, and knowledge. I would also like to thank Dr. Luis G Sanchez Giraldo and Md Nasir Uddin Laskar for helping me understand what I was doing and for allowing me to build on their work. A special thanks to David W Grossman for helping me get started with my work. I would also like to thank the University of Miami Department of Computer Science for allowing me to temporarily commandeer their GPUs. Finally, I want to thank the National Science Foundation for funding my work and allowing me and other curious college students to be part of this REU experience.

References

[1] C. M. Ziemba, J. Freeman, J. A. Movshon, and E. P. Simoncelli, "Selectivity and tolerance for visual texture in macaque V2," *Proceedings of National Academy of Science (PNAS)*, vol.113, no. 22, 2016.

[2] J. Freeman, C. M. Ziemba, D. J. Heeger, E. P. Simoncelli, and J. A. Movshon, "A functional and perceptual signature of the second visual area in primates," *Nature Neuroscience*, vol. 16, no.7, 2013.

[3] N. Kriegeskorte *Cold Spring Harbor Labs*, 2015.

[4] A. Krizhevsky, I. Sutskever, G. E. Hinton. "ImageNet classification with deep convolutional nueral networks," in *Advances in neural information processing systems (NIPS)*, 2012.

[5] M. Zeiler, and R. Fergus: "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision (ECCV)*, 2014.

[6] D. J. Dong, et. al.: *Proc. CVPR*, 2009

[7] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015

[8] D. L. Yamins, H. Hong, C.F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized heirarchical models predict neural responses in higher visual cortex," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, 2014.