UNIVERSITY OF MIAMI
DEPARTMENT of
COMPUTER SCIENCE

The Effect of Divisive Normalization in Deep Convolutional Neural Networks on Figure Ground Assignment

Nicholas Cimaszewski[1], Luis Gonzalo Sánchez Giraldo[2], Odelia Schwartz[2]
[1]Goergen Institute of Data Science, University of Rochester, [2]Department of Computer Science, University of Miami

Background

CNN's have proven to be some of the most effective computer vision techniques. Due to the many parallels found between these systems and biological vision, attention has been paid to designing CNN's after the visual stream in search of task based improvement as well as more accurate neuroscience models.

In the research presented here, we implement divisive normalization, a canonical neural computation that has been shown in areas such as primary and secondary visual cortex (V1 and V2), at a point in a CNN comparable to these areas and test whether it improves performance on figure-ground assignment.

Flexible Surround Normalization

Divisive normalization operates by scaling each unit's output by the sum of a pool of surrounding units and adjacent channels. In *flexible surround normalization*, the method we used, the influence of the spatial surround depends on the statistical dependence between the center and surround response.

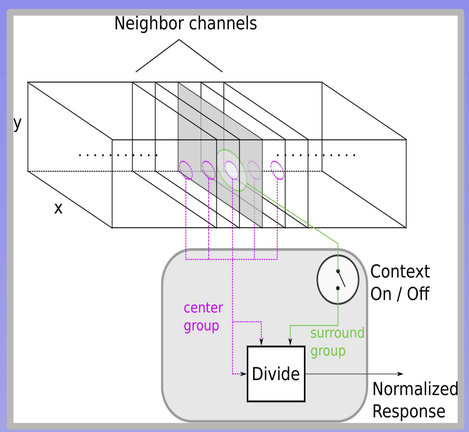


Figure 1
Flexible surround normalization.

Taken from Gonzalo Sánchez Giraldo and Schwartz (2018).

Figure-Ground Assignment

Figure-ground assignment was measured by running small patches of images with figure-ground labelings from the Berkeley Segmentation Dataset (BSDS) partway through a pretrained model of AlexNet, the landmark 2012 image classification network. This experiment's design inherits much from Coen-Cagli & Schwartz (2013), which applied divisive normalization to a V1 model and measured performance of the same task on the same dataset.

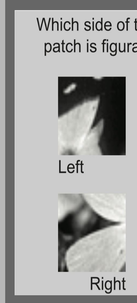


Figure 2
From Coen-Cagli and Schwartz (2013).

Implementation and Network Architecture

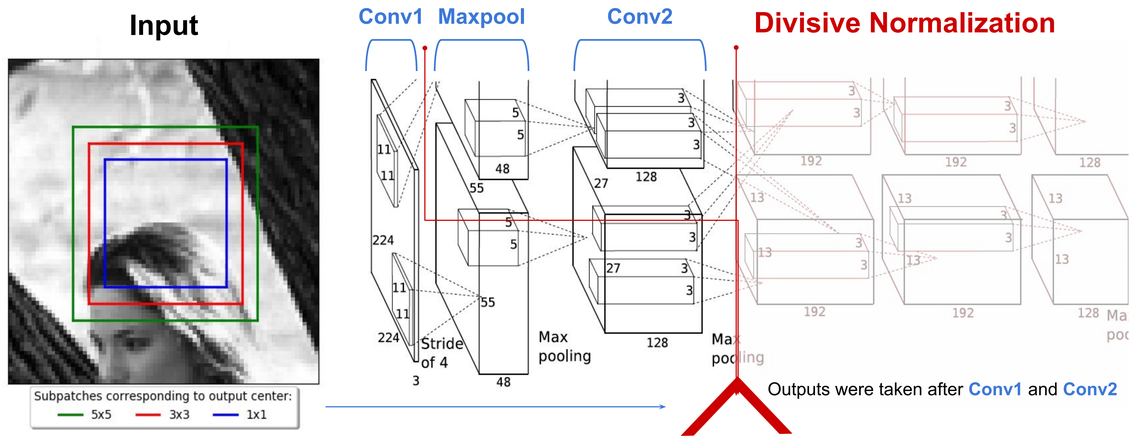
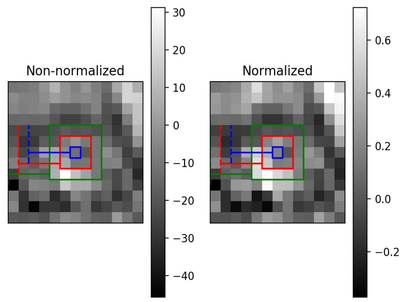


Figure 3
The architecture of AlexNet is depicted above. An example patch is shown on the left ("Input"). The highlighted regions of the patch are the pixels corresponding to the respective regions of the output maps (right). The surround regions needed for normalizing each center square in the output is shown by the line extending out from the middle. Everything past the red line was omitted - this is the last point at which we extracted the output and applied normalization.



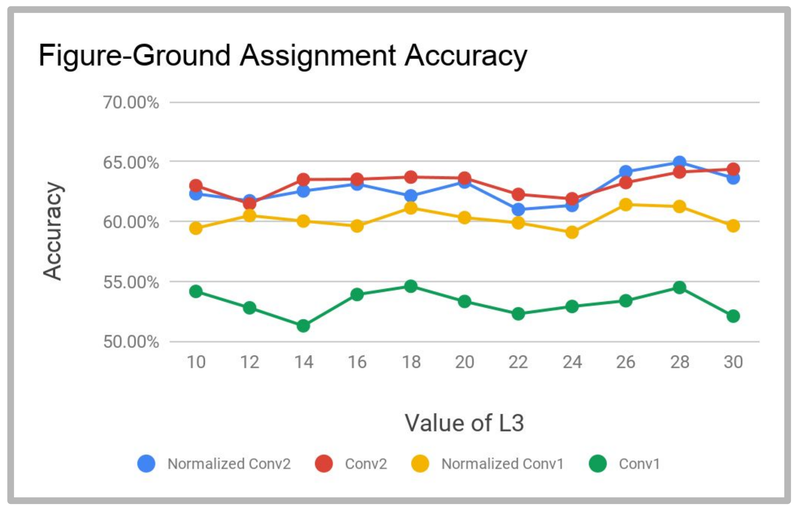


Figure 4
Logistic regression with 10-fold cross validation was performed on the output maps and their accuracies in predicting which side of the input patch was the background was recorded. Results for the different output maps are shown on left.
L3 is a parameter of the image patches generated that specified the minimum distance from junctions of 3 or more contours to the contour point at the center of the patch. Although we predicted a larger L3 would result in less cluttered patches and improve assignment accuracy, it was not a factor.

Results

Accuracy maxed out at around 64% accuracy, and there seemed to be no difference made by the normalization applied after Conv2. Most notably, it did not achieve the 76% accuracy reached by Coen-Cagli and Schwartz (2013). The principal difference between our methodology and theirs is that they used a V1 model with filters statistically designed for efficient coding, whereas we adapted the filters of a CNN trained to perform object classification. The fact that AlexNet could not match the success of a model with fewer layers suggests that filters more attuned to higher level tasks like object classification are different from those better suited for mid-level tasks like figure-ground assignment.

One similarity we do see to Coen-Cagli and Schwartz's results is that after Conv1, normalized response predicted figure-ground ~5% better, similar to the margin of improvement due to normalization after their V1 model. However, this margin disappears in our Conv2 responses, suggesting that the roles of Conv2 and divisive normalization may be redundant in a CNN.

Next steps include adding normalization at other points in the architecture and even at multiple layers and applying this methodology to other mid-level tasks. While the lack of improvement after Conv2 was surprising, it implies that divisive normalization is not a meaningful computation for figure-ground assignment at this point.

Acknowledgements

My deepest thanks to Dr. Sánchez Giraldo and Dr. Schwartz for all of their guidance through the many challenges we encountered. Thank you as well to Dr. Ruben Coen-Cagli for help in generating the dataset of patches. Thank you to Dr. Rosenberg for organizing our program and leading many workshops.

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1659144.

References

Coen-Cagli, R., & Schwartz, O. (2013). The impact on midlevel vision of statistically optimal divisive normalization in V1. Journal of Vision, 13(8), 13-13. doi:10.1167/13.8.13

Carandini, M., & Heeger, D. J. (2011). Normalization as a canonical neural computation. Nature Reviews Neuroscience, 13(1), 51-62. doi:10.1038/nrn3136

Williford, J. R., & Heydt, R. V. (2016). Figure-Ground Organization in Visual Cortex for Natural Scenes. Eneuro, 3(6). doi:10.1523/eneuro.0127-16.2016

Gonzalo Sánchez Giraldo, L. & Schwartz, O. (2018). "Integrating Flexible Normalization into Mid-Level Representations of Deep Convolutional Neural Networks." arXiv e-prints. arXiv:1806.01823

C. Fowlkes, D. Martin, J. Malik. "Local Figure/Ground Cues are Valid for Natural Images" Journal of Vision, 7(8):2, 1-9.