Figure-ground representation in deep neural networks

Brian Hu*, Salman Khan†, Ernst Niebur‡, Bryan Tripp†
*Allen Institute for Brain Science
Seattle, WA 98109
†Centre for Theoretical Neuroscience
University of Waterloo
Waterloo, Ontario, Canada
‡Zanvyl Krieger Mind/Brain Institute
Johns Hopkins University
Baltimore, Maryland 21218

***Abstract*—Deep neural networks achieve state-of-the-art performance on many image segmentation tasks. However, the nature of the learned representations used by these networks is unclear. Biological brains solve this task very efficiently and seemingly effortlessly. Neurophysiological recordings have begun to elucidate the underlying neural mechanisms of image segmentation. In particular, it has been proposed that border ownership selectivity (BOS) is the first step in this process in the brain. BOS is a property of an orientation selective neuron to differentially respond to an object contour dependent on the location of the foreground object (figure). We explored whether deep neural networks use representations close to those of biological brains, in particular whether they explicitly represent BOS. We therefore developed a suite of *in-silico* experiments to test for BOS, similar to experiments that have been used to probe primate BOS. We tested two deep neural networks trained for scene segmentation tasks (DOC [1] and Mask R-CNN [2]), as well as one network trained for object recognition (ResNet-50 [3]). Units in ResNet-50 predominantly showed contrast tuning. Units in Mask R-CNN responded weakly to the test stimuli. In the DOC network, we found that units in earlier layers of the network showed stronger contrast tuning, while units in deeper layers of the network showed increasing BOS. In primate brains, contrast tuning seems wide-spread in extrastriate areas while BOS is most common in intermediate area V2 where the prevalence of BOS neurons exceeds that of earlier (V1) and later (V4) areas. We also found that the DOC network, which was trained on natural images, did not generalize well to the simple stimuli typically used in experiments. This differs from findings in biological brains where responses to simple stimuli are stronger than to complex natural scenes. Our methods are general and can also be applied to other deep neural networks and tasks.**

I. Introduction

Deep learning enables machines to learn representations useful for a variety of real-world tasks in an end-to-end manner, with applications to object recognition, object detection, and scene segmentation [4], [5]. Here, we focus on the problem of instance segmentation, where each separate object instance within an image has to be segmented on a pixel-wise basis. For this specific problem, various datasets (e.g. COCO [6] and Cityscapes [7]) and models, e.g. Deep Occlusion Estimation (DOC) [1] and Mask R-CNN [2] have been proposed. As these approaches are mostly task-driven, researchers typically optimize model performance with respect to an error metric (e.g. mean average precision of the intersection over union on the instance segmentation task). Much less attention is usually paid to the learned representations that result from training on a specific task, although a related line of research on model interpretability has provided some useful insights [8], [9].

From the neuroscience perspective, the brain is able to solve visual tasks very efficiently, without the need for the large numbers of labeled examples typically used to train deep neural networks (for a discussion on the connections between neuroscience and deep learning, see [10]). For the specific problem of figure-ground segmentation, Zhou et al [11] showed that individual neurons in primate visual cortex implement border ownership coding in their firing rates. Depending on which side an object is located relative to its receptive field, a border ownership selective neuron will respond with different firing rates. The difference in firing rates for when an object is located on the neuron's preferred side (which has a higher firing rate) versus when it is located on its non-preferred side (with a lower firing rate) can then be used to infer figure-ground relationships. Border ownership coding has been observed for a large battery of stimuli, including both simple stimuli [11]–[15] and complex natural scenes [16], [17]. Several mechanisms have been proposed to explain this phenomenon, based on local (intra-areal) processing [18], feed-forward architectures [19], [20], or the interaction of bottom-up and top-down processes [21].

Here, we propose to use deep neural networks trained on scene segmentation tasks and study their learned representations in a manner similar to conventional neurophysiological experiments by "recording" from units within the network (similar to [22]). Our *in-silico* experiments allow us to directly compare the representations learned by neural networks and those which have been observed in the brain.

II. Methods

A. Networks

We performed *in-silico* experiments on two networks that have been pretrained to perform figure-ground segmentation

in different ways [1], [2]. One of these networks (DOC, ref. [1]) performs both contour detection and figure-ground segmentation using a two-stream architecture. The model was trained on natural images with manually annotated object contour and figure-ground labels. For a given image, the model produces two outputs– a contour map and a figure-ground orientation map, which are then fused together to produce a pixel-wise segmentation of objects within the image. For a more in-depth overview of the model, we refer the reader to the original paper [1]. The source code for this project is publicly available at: https://github.com/pengwangucla/DOC. We first converted the original model from Caffe to Tensorflow using the Microsoft MMdnn toolbox. Our subsequent analyses were performed using the Tensorflow framework.

The second network (Mask R-CNN, ref [2]) has several parts that collectively perform instance segmentation. Different sub-networks propose object bounding boxes, label bounded objects by object category, and propose masks over pixels that correspond to each object. We focused on the mask sub-network. We used the pretrained Caffe2 model provided by [2]. As a baseline, we also examined a standard object-recognition network, ResNet-50 [3]. We used the pre-trained implementation that is packaged with Keras. Our experiment and analysis code can be found online at: https://github.com/brianhhu/DOC-tf.

B. Stimuli

The input to the network consisted of stimuli that were similar to those used in neurophysiological studies in extrastriate cortex of awake behaving monkeys. The simplest stimulus used to characterize BOS responses in experimental neurophysiology is a square foreground object in front of a uniform background, see ref [11], Fig. 2. It was found that responses are essentially invariant when the size of the stimulus is changed, *ibid* Figs 5, 9. Several variations of the square stimulus were used, *e.g.* overlapping squares (*ibid*, Fig. 3), a square composed of "Cornsweet contours," ref [12] Fig. 1, or of non-connected segments, ref [12] Fig. 3. Other geometrical shapes were used to demonstrate the general nature of the phenomenon, including a "C-shape," ref [11] Fig. 2 and variations of it, ref [21] Fig. 9. The standard stimulus against which others are typically compared in many of these situations, as well as in complex natural scenes [17] is, however, the square which is therefore used in our computational experiments.

We studied the response properties of units across layers and feature maps of the networks. We always "recorded" from the center unit within a given feature map, and we centered and rotated the edge of the square stimulus in order to reliably drive each unit. We adapted the color and orientation of the square stimulus to the preferred color and orientation of the unit under study using the same methods outlined in the original experiments [11]. Briefly, we first used a set of bar stimuli centered within the image and varied the length, width, color, and orientation of the bar in order to find the preferred stimulus of the unit under study.

The standard square test involves four stimulus conditions, including all combinations of two binary factors: side-of-figure of the square, and contrast of the figure edge (examples shown in Fig. 1, left four panels). Except where noted, we used squares that were 1/4 the width of the full stimulus image (we used square images 400 pixels wide for the DOC and Mask R-CNN networks, and 256 pixels wide for ResNet). To determine whether a unit had stronger contrast tuning or border ownership tuning, we calculated a "contrast score" and a "border ownership score," respectively. Given the responses R_A, R_B, R_C, and R_D to stimuli A, B, C, and D in the square test (see [11] and Fig. 1), the contrast score was

$$S_C = |(R_A + R_B)/2 - (R_C + R_D)/2|/\bar{R} * 100, \quad (1)$$

where $\bar{R}$ is the mean of R_A, R_B, R_C, and R_D. This score ranged from 0 to 200. Similarly, the border ownership score was

$$S_B = |(R_A + R_C)/2 - (R_B + R_D)/2|/\bar{R} * 100. \quad (2)$$

These scores are loosely related to the reliability scores that were used to characterize response strength in [11].

Contrast tuning is a local measure that captures whether units code for the contrast polarity of a figure edge, independent of where the square stimulus is located. On the other hand, border ownership coding is a global measure that captures whether units code for side-of-figure, independent of the contrast polarity of the figure and background. We quantify the degree of contrast or border ownership tuning using scatter plots of these two measures for units across the feature maps within a given layer of the network. Points falling on the unity line indicate equal contrast and border ownership tuning, while points below/above the unity line indicate preference for either contrast or border ownership tuning.

Additionally, we found optimal stimuli for certain units, using standard methods in neural network analysis. Starting with a random image, we used backpropagation to find the gradient of a given unit's activation, with respect to the input image, and used gradient ascent to find an image that strongly activated the unit [8], [9]. High-frequency noisy gradients are suppressed using Laplacian pyramid gradient normalization [23]. Optimal stimuli for two example neurons are shown in Figure 8.

III. Results

We focus mainly on the Deep Occlusion Estimation (DOC) network, which had the strongest border ownership coding.

Figure 1 shows an example set of square stimuli for the standard square test, along with corresponding outputs of the orientation and boundary detection streams of the DOC network. The network failed to mark the boundaries of the squares used in the standard test, in contrast with its good performance with natural images, e.g. Figure 2.

Figure 3 shows distributions of contrast preference and border preference scores, defined respectively in eqs 1 and 2, for every third ReLU layer in the DOC network's orientation stream. There were strong contrast-selective responses

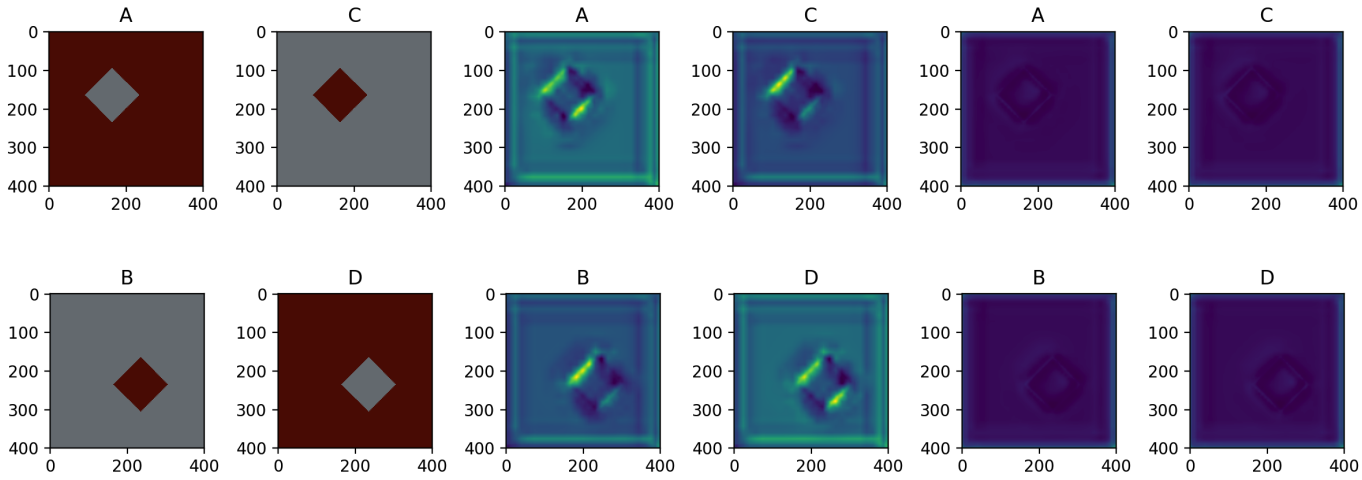

Fig. 1. Example output of the pretrained DOC network on the standard square test. The left panels show the original input images, the center panels show the outputs of the figure-ground orientation stream of the network, and the right panels show the outputs of the contour detection stream of the network. Each group of four figures shows the different side-of-figure and contrast conditions that were tested (i.e. left and right differ by contrast, up and down differ by side-of-figure). We "recorded" from units located in the exact center of the image. Importantly, the local image content within the receptive field of these units was the same between side-of-figure conditions (compare A and B or C and D).

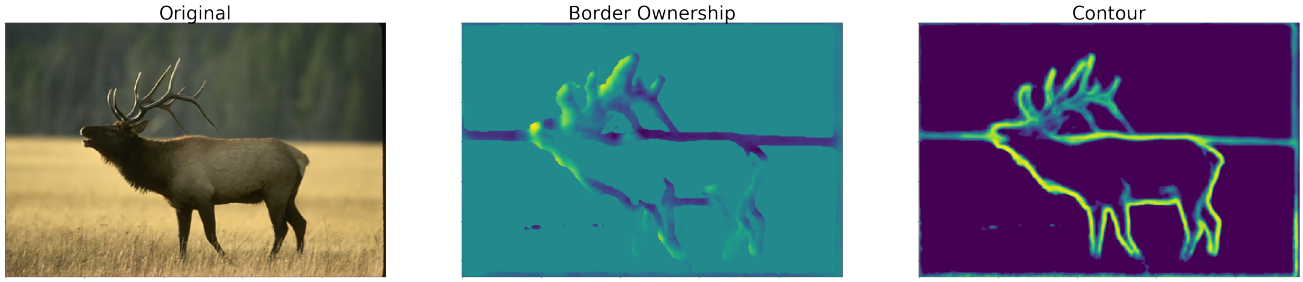

Fig. 2. Example output of the pre-trained DOC network. The left panel shows the original input image, the center panel shows the output of the contour detection stream of the network, and the right panel shows the output of the figure-ground orientation stream of the network. The example image is from the Berkeley Segmentation Data Set [24].

throughout the network. Border ownership scores were much smaller than contrast scores in earlier layers, but comparable to contrast scores in later layers. The last layer (relu5_3) contained a number of units with minimal contrast sensitivity and strong border selectivity (lower-right corner of scatter-plot in Figure 3). Overall, the distributions of contrast and border ownership responses were similar in this layer. This qualitatively resembles the pattern of responses to similar stimuli in V2 [11]. Because V2 responses are variable across repeated trials, Zhou *et al.* [11] characterized coding of contrast and ownership using an ANOVA. To compare more directly with these results, we added a Poisson model of spike-rate variability to the relu5_3 responses, and performed the same analysis. We found that 43% of units had statistically significant representations of both contrast and ownership ($\alpha = 0.01$), 18% significantly encoded contrast only, and 15% significantly encoded ownership only. These proportions are similar to those in V2 (44%, 22%, and 15%, respectively) [11]. However, the model did not show size invariance at any level which is observed in the biological data. As can be seen in Figure 4, earlier layers have a strong preference for the standard square (100×100 pixels) while later layers prefer the larger square (150×150). Figure 5 summarizes the relative strength of contrast and border coding throughout the network.

Figure 6 plots the relative strength of contrast and border tuning in two other networks: ResNet and Mask R-CNN. There were strong contrast responses and weak border responses throughout the ResNet. The vast majority of units in the Mask R-CNN network had weak responses to the square stimuli.

In addition to square figures, [11] also experimented with C-shaped figures, and stimuli with two overlapping figures. Each of these groups of stimuli had the same local contrast, at the centre, as the group of square stimuli. [11] found that somewhat less than half of cells with significant border preference in response to square stimuli also had a significant preference in response to the more complex stimuli. We tested responses of layer relu5_3 of the DOC network's orientation stream to these additional stimuli. Figure 7 shows border-preference scores with C-shaped (left) and overlapping-figures (right). Scores in response to the more complex stimuli (verti-

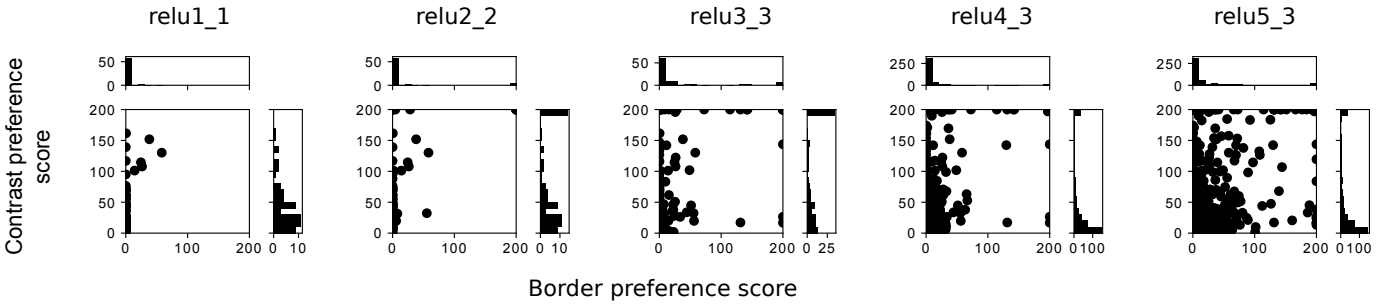

Fig. 3. Comparison of contrast and border ownership tuning in every third layer of the figure-ground orientation stream of the DOC network. The border preference score is defined in eq 2 and contrast reference score in eq 1. The first convolutional layer (relu1_1, left) largely shows contrast tuning. Later layers show progressively stronger border ownership tuning. The final convolutional layer (relu5_3) shows a mix of both contrast and border ownership responses. We find similar results in the contour-detection stream of the DOC network (not shown).

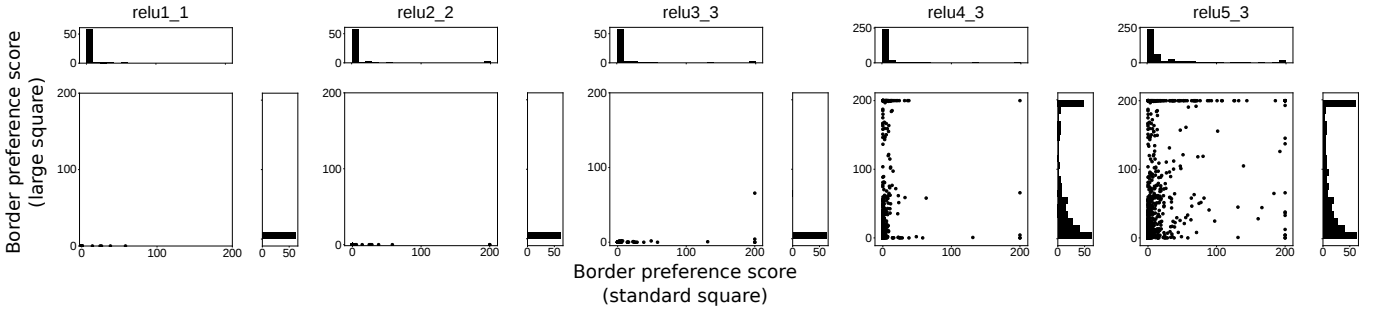

Fig. 4. Lack of size invariance of border ownership tuning in the figure-ground orientation stream of the DOC network. Plotted is the border ownership score (eq 2) of the large *vs* that of the standard square. Earlier layers (relu1_1, relu2_2, and relu3_3) show preference for the standard square size (100x100 pixels). Later layers show preference for a larger square size (150x150 pixels). We find similar results in the contour-detection stream of the DOC network (not shown).

cal axes) are plotted against those in response to square stimuli (horizontal axes) for the same units. In contrast with [11], border preferences were at least as strong with the complex stimuli as with the square stimuli. The same was true for layer relu5_3 of the boundary stream (not shown).

Among cells with significant border preferences in response to C shapes and overlapping figures, [11] found that nearly all such cells preferred complex stimuli on the same side (e.g. left or right) as they preferred square stimuli. We tested whether units with strong border preferences (scores >50) with both squares and another stimulus group preferred the same side for each kind of stimulus. Similar to [11], 16/18 such units in boundary stream layer relu5_3 preferred C-shape and square stimuli on the same side, and 7/8 preferred overlapping and square stimuli on the same side. In contrast, side preferences were shared across stimuli less often in orientation-stream layer relu5_3. Only 39/58 units had matching side preferences for square and C-shaped stimuli, and only 17/36 had matching side preferences for square and overlapping stimuli.

To better understand the mechanism of strong border ownership responses in later layers of the DOC network, we found optimal stimuli for some of these units. We focused on units in layer relu5_3 of the orientation stream, and selected units with a strong border or contrast bias, specifically $|S_C - S_B| > 150$. Figure 8 shows one example of optimal stimuli for a unit with a strong border response, and for another one with a strong contrast response. Importantly, these units' receptive fields are larger than the square stimuli, allowing them to differentiate stimuli based on the surrounding region. To see whether the optimal stimuli predicted the border and contrast responses, we performed a simple approximation of the square experiment. We modelled each unit's response to a square stimulus I^{sq} as $\max(0, \sum_{i,j,k} I^{sq}_{i,j,k} I^{opt}_{i,j,k})$, where I^{opt} is the optimal stimulus (a variant without the Laplacian pyramid method), and the subscript i, j, k refers to the $(i, j)^{th}$ pixel in the k^{th} channel. Two of five units with strong border responses also had substantial border responses in this simplified model, compared to zero of seven units with strong contrast responses. This suggests that similarity of various square stimuli to these units' optimal stimuli may partly account for the results, despite the many nonlinearities in the network.

IV. Conclusion

We developed a set of tools that allowed us to perform an *in-silico* version of neurophysiological experiments testing for border ownership coding in deep neural networks, which allowed for direct comparison of neural responses and responses from units within the network. Our results show that while deep neural networks learn representations that enable them to solve difficult visual tasks such as instance segmentation, these representations can differ greatly from those that have been experimentally observed in the brain. We found that although

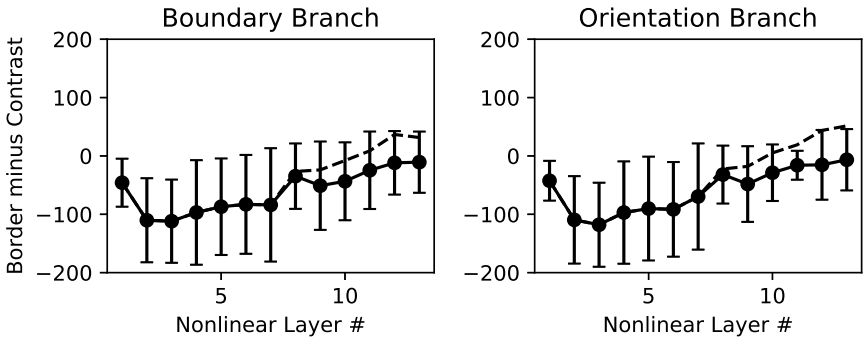

Fig. 5. Summary of differences in the strength of border and contrast tuning across the DOC network. Left: mean +/- standard deviation of border score minus contrast scores in each layer of the orientation stream. Right: mean +/- standard deviation of border minus contrast scores in each layer of the boundary stream. Lower values indicate stronger contrast than border tuning. The dashed lines indicate the means with smaller square stimuli (50×50 pixels instead of 100×100).

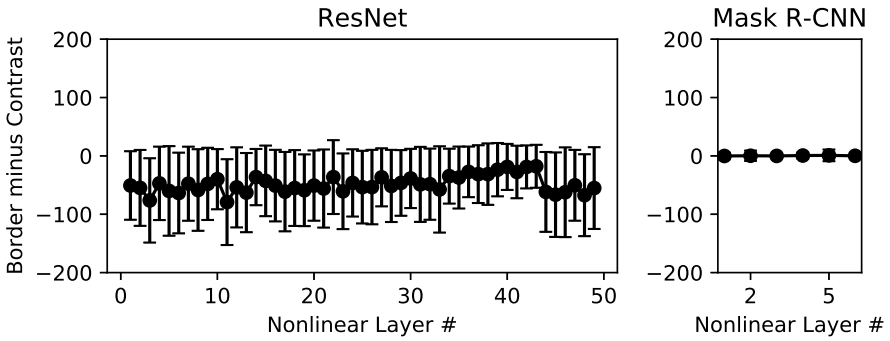

Fig. 6. Summary of differences in the strength of border and contrast tuning in other networks (conventions as in Figure 5). Left: Contrast responses dominate throughout ResNet, although only weakly in nonlinear layers 39-43. Right: The mask branch of Mask R-CNN responded weakly to the standard square stimuli.

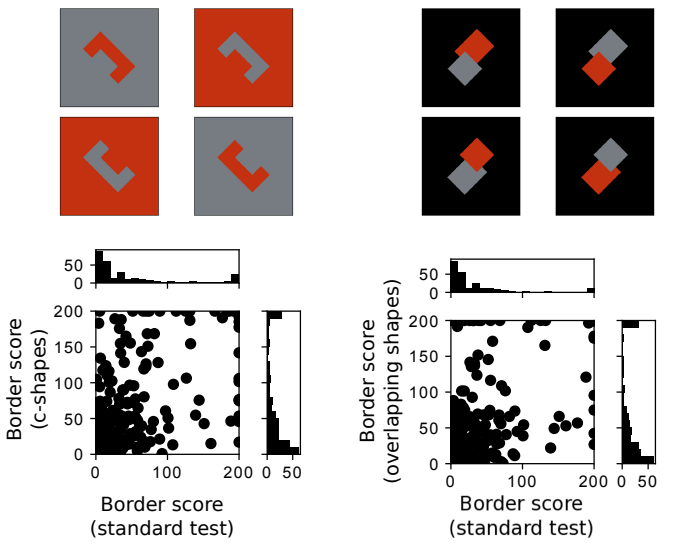

Fig. 7. Border-preference scores of layer relu5_3 of the DOC network's orientation stream, with the C-shape and overlapping-figures stimuli used in [11] (top panels). The bottom panels show border-preference scores of responses to these stimuli (vertical axes) against the border-preference scores of responses of same units to the standard square stimuli (horizontal axes).

higher layers showed stronger border ownership tuning, corresponding to a global image property related to figure-ground segmentation, a large number of neurons in these layers were still sensitive to contrast, a largely local image cue. We also found that contour and figure-ground responses were weaker to the artificial stimuli we used, compared to the natural stimuli that the model was trained on, suggesting a possible generalization gap. We note that in biological networks, the opposite was found: the BOS signal for the standard square was typically several times larger than that for foreground objects in complex natural scenes [17]. It seems quite possible that the cues used for solving the segmentation task, as well as the mechanisms by which they are exploited and their internal representations are fundamentally different between the two systems. This is because their architectures differ in a principled way: while most convolutional neural networks, including all three networks which we study as an example here, have a feedforward architecture, biological brains are highly recurrent. Given the primate visual system's general tendency of increasing receptive field sizes in more central compared to more peripheral brain areas, feedback from higher areas provides access to a large context which is likely used to solve the segmentation problem. This context information

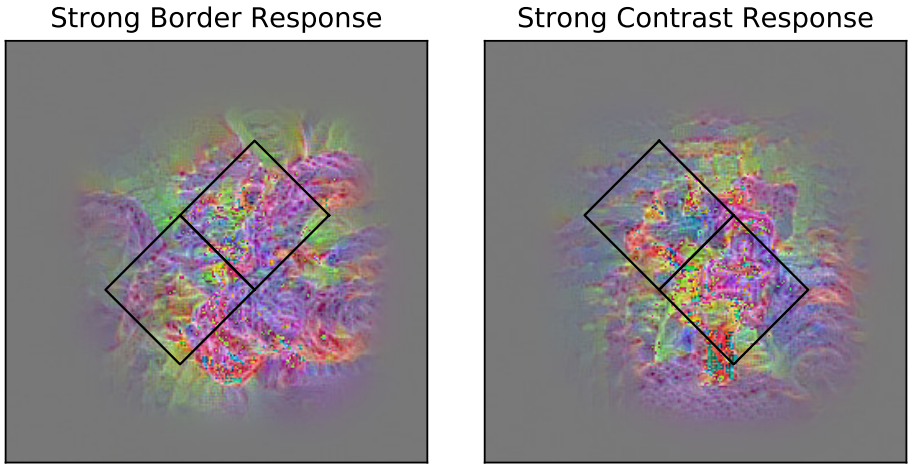

Fig. 8. Optimal stimuli for example units with strong border (left) and contrast (right) responses. Both units are from layer relu5_3 of the DOC network's orientation stream. The black outlines indicate the positions and orientations of the square stimuli.

is not available in early layers of a feedforward network where receptive fields are by design small. Strictly feedforward networks therefore may need to employ different strategies to solve the problem, or at least come as close to a solution as possible. Future work will include the study of additional networks with different architectures, and the design of new network architectures better constrained by the figure-ground representations observed in biology.

Acknowledgment

Supported by the National Science Foundation through grant 1835202 and by NIH through R01DA040990 and R01EY027544.

References

[1] P. Wang and A. Yuille, "Doc: Deep occlusion estimation from a single image," in *European Conference on Computer Vision*. Springer, 2016, pp. 545–561.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, 2017.

[5] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, 2017, https://distill.pub/2017/feature-visualization.

[9] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, "The building blocks of interpretability," *Distill*, 2018, https://distill.pub/2018/building-blocks.

[10] A. H. Marblestone, G. Wayne, and K. P. Kording, "Toward an integration of deep learning and neuroscience," *Frontiers in computational neuroscience*, vol. 10, p. 94, 2016.

[11] H. Zhou, H. S. Friedman, and R. von der Heydt, "Coding of border ownership in monkey visual cortex," *J. Neurosci.*, vol. 20, no. 17, pp. 6594–6611, 2000, pMID: 10964965.

[12] N. Zhang and R. von der Heydt, "Analysis of the context integration mechanisms underlying figure–ground organization in the visual cortex," *The Journal of Neuroscience*, vol. 30, no. 19, pp. 6482–6496, 2010, pMC2910339.

[13] F. T. Qiu and R. von der Heydt, "Neural representation of transparent overlay," *Nat. Neurosci.*, vol. 10, no. 3, pp. 283–284, 2007.

[14] P. O'Herron and R. von der Heydt, "Remapping of Border Ownership in the Visual Cortex," *The Journal of Neuroscience*, vol. 33, no. 5, pp. 1964–1974, 2013, pMID: 23365235 [PubMed - in process].

[15] A. B. Martin and R. von der Heydt, "Spike Synchrony Reveals Emergence of Proto-Objects in Visual Cortex," *The Journal of Neuroscience*, vol. 35, no. 17, pp. 6860–6870, 2015.

[16] J. R. Williford and R. von der Heydt, "Border-ownership coding," *Scholarpedia*, vol. 8, no. 10, p. 30040, 2013.

[17] ——, "Early Visual Cortex Assigns Border Ownership in Natural Scenes According to Image Context," *Journal of Vision*, vol. 14, no. 10, pp. 588–588, 2014.

[18] L. Zhaoping, "Border ownership from intracortical interactions in visual area V2," *Neuron*, vol. 47, pp. 143–153, 2005, pMID: 15996554.

[19] H. Nishimura and K. Sakai, "The computational model for border-ownership determination consisting of surrounding suppression and facilitation in early vision," *Neurocomputing*, vol. 65, pp. 77–83, 2005.

[20] A. F. Russell, S. Mihalas, R. von der Heydt, E. Niebur, and R. Etienne-Cummings, "A model of proto-object based saliency," *Vision Research*, vol. 94, pp. 1–15, 2014.

[21] E. Craft, H. Schütze, E. Niebur, and R. von der Heydt, "A neural model of figure-ground organization," *Journal of Neurophysiology*, vol. 97, no. 6, pp. 4310–26, 2007, pMID: 17442769.

[22] B. P. Tripp, "Similarities and differences between stimulus tuning in the inferotemporal visual cortex and convolutional networks," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 3551–3560.

[23] A. Mordvintsev, C. Olah, and M. Tyka, "Inceptionism: Going deeper into neural networks," 2015.

[24] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," in *Proc. 8th Int'l Conf. Computer Vision*, vol. 2, July 2001, pp. 416–423.