Evidence for Visual Representation of Numerosity in Natural Scenes

Maggie Mae Mell (mell@musc.edu)
Ghislain St-Yves (stayves@musc.edu)
Department of Neuroscience, 173 Ashley Avenue
Charleston, SC 29425 USA

Emily J. Allen (prac0010@umn.edu)
Department of Psychology, 75 E River Rd
Minneapolis, MN 55455 USA

Yihan Wu (wuxx0468@umn.edu)
Kendrick Kay (kay@umn.edu)
Department of Radiology, 2021 6th Street SE
Minneapolis, MN 55455 USA

Thomas Naselaris (tnaselar@musc.edu)
Department of Neuroscience, 173 Ashley Avenue
Charleston, SC 29425 USA

Abstract

In visual cortex of human and non-human primates, high-level visual areas near intraparietal sulcus have been shown to explicitly encode the number of objects in visual displays. To date, evidence for this numerosity code has come from experiments that use simple dot-like visual stimuli, raising the question of whether the numerosity code persists during perception of natural scenes. Here, we assessed evidence for a numerosity code in high-resolution fMRI measurements of responses to thousands of natural scenes in 3 human subjects. We constructed an encoding model that predicted voxelwise responses as a function of local object counts in each natural scene. Our model was able to accurately predict voxelwise activity in visual cortex. To test if local object counts were acting as a proxy for simple low-level image features, we constructed voxelwise encoding models based on Gabor wavelet filtering of the natural scenes. For voxels in anterior visual cortex, the numerosity encoding model generated more accurate predictions than the Gabor model. These results offer preliminary evidence for a numerosity code in anterior visual cortex during natural scene stimulation.

Keywords: fMRI; encoding models; natural scenes; numerosity

Introduction

Humans and non-human primates are able to quickly estimate the number of objects in visual space. Previous studies have suggested that this ability is linked to maps in parietal cortex that are tuned to the number of objects in simple dot-like displays (Harvey, Klein, Petridou, & Dumoulin, 2013; Nieder & Miller, 2004; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004; Tudusciuc & Nieder, 2007). However, in the natural world we are rarely presented with cleanly segmented and separated objects. In contrast, natural scenes contain many different objects with varied shapes, sizes, occlusion, etc. It is not known if the numerosity representation found in previous studies persists when viewing complex natural stimuli. In this paper we developed a numerosity-based voxelwise encoding model to explore the representation of numerosity in the human brain in response to natural scene stimulation. We present preliminary evidence for a numerosity representation that appears to be distinct from low-level, wavelet-like features, but is subsumed by more complex feature representation in a performance-optimized deep neural network.

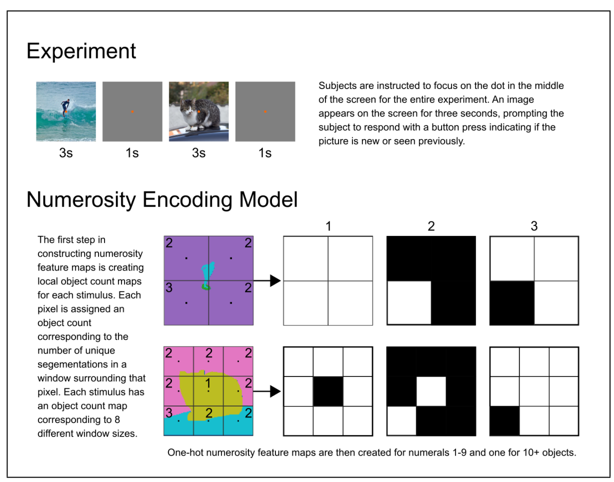

Figure 1: Experiment. fMRI BOLD responses were recorded in response to thousands of natural scene images. Model. The Numerosity Encoding Model was constructed by calculcating local object counts for every natural scene image. Object counts were transformed into one-hot maps for numerals 1-9 and 10+.

Methods

Data

High-resolution whole-brain fMRI BOLD measurements were obtained from human subjects in response to natural scene images (Figure 1 top). Each subject was presented with 9,000 unique images and 1,000 images shared among all subjects, totalling 10,000 images each. Each image was shown 3 times, non-sequentially, for a total of 30,000 stimulus trials presented over 40 separate runs. Images are displayed for 3 seconds followed by a 1 second interstimulus interval. Subjects were instructed to focus on a center fixation dot and respond with a button press to indicate if the image is new or repeated. Data from N=3 subjects are presented in this paper. Further data collection is ongoing.

Encoding Models

Three independent encoding models were fit to voxel responses. All models were trained using a feature-weighted receptive field (fwRF) ridge regression method, previously developed by our lab (St-Yves & Naselaris, 2018). Briefly the fwRF is a voxel-wise encoding model that estimates receptive field location and size from stimulus generated feature-maps. The fwRF generates predictions of brain activity in response to a visual stimulus based upon feature weights and a feature pooling field. The feature pooling field indicates the region in visual space a voxel's activity is most driven by. The same feature maps are used across voxels, however the weights assigned to each feature will vary and indicate features encoded in the activity of each voxel. Values for the location and radius of the feature-pooling field, i.e. the fwRF center and size, as well as the feature weights were estimated using ridge regression. Three distinct sets of feature maps were used to generate three models of brain activity.

Numerosity Encoding Model To test for numerosity coding we constructed an encoding model that predicted voxel-wise responses as a function of local object counts in each image (Figure 1 bottom). Feature maps for this model were created based on panoptic scene segmentations of publically available images from the COCO dataset (Lin et al., 2014). Segmentation images were cropped and down-sampled to 128x128 pixels. Object counts for each pixel were assigned based upon how many unique segmentations appear in a window around that pixel. 8 window sizes were used, 16x16 to 128x128 in steps of 16. One-hot maps were then created for numerals 1-9 and 10+, giving 80 feature maps. One-hot maps are simply an indication of whether the window around that pixel contained the number of objects specific to that map. For instance, if a 16x16 window around a pixel contained 3 unique objects, that pixel would have a 1 in the feature map for numeral 3 and a 0 in all other one-hot maps for that window size.

Alternate Encoding Model For comparison we constructed two alternate encoding models based on Gabor-wavelet feature maps (St-Yves & Naselaris, 2018), and feature maps extracted from a deep neural network (Krizhevsky, Sutskever, & Hinton, 2012), respectively.

Prediction accuracy and cross-validation

All encoding models were trained on ~80% of responses from each subject. Approximately 20% of the training set was held-out for optimization of ridge hyper-parameter as well as selection of fwRF location and sizes. The remaining ~20% were used for cross-validation. Prediction accuracy is the Pearson correlation between model predictions and measured voxel responses.

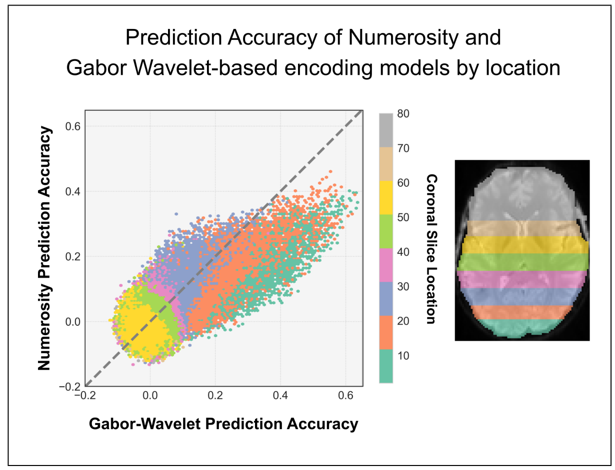

Figure 2: Prediction Accuracy Model Comparison. The joint distribution of prediction accuracy (Pearson correlation between predicted and measured brain activity from 3 subjects for Gabor wavelet-based encoding model (x-axis) and Numerosity based encoding model (y-axis). Voxels are binned into hexagons and colored by the average coronal slice location for all voxels in that bin. An example axial slice on the left shows approximate slice demarcations for each color category (surface reconstructions and ROIs are not yet available for this preliminary dataset). The Gabor-wavelet encoding model makes more accurate predictions than the Numerosity model in posterior sections of the brain, especially early visual areas. In anterior visual (slice locations 20-30, which includes parietal visual areas) areas the Numerosity encoding model makes more accurate predictions.

Results

In natural scenes numerosity is likely to be correlated with, but not entirely determined by, simple low-level attributes (e.g. contrast, power variation across spatial frequency bands) of images. To determine where or if the encoding of numerosity and low-level features is disentangled in visual cortex we compared the prediction accuracy of the Numerosity and Gabor-wavelet encoding models. For many voxels in anterior visual cortex the Numerosity encoding model was the more accurate predictor of brain activity (Figure 2). This result suggests that anterior visual cortex maintains a representation of the local number of objects in natural scenes, and that this representation cannot be entirely accounted for by simple low-level image attributes.

A recent study (DeWind, 2019) suggests that numerosity representations can arise "spontaneously" in feature maps of deep convolutional neural networks (DCNN) trained to categorize objects (Krizhevsky et al., 2012). This result suggests that the Numerosity encoding model should exhibit no advantage in prediction accuracy over an encoding model based on the feature maps of an optimized DCNN. Indeed, the DCNN-based encoding model outperforms the Numerosity encoding model for all voxels (Figure 3). This result clearly demonstrates that a representation of the local number of objects in natural scenes is only a subset of the representations maintained in all visual areas.

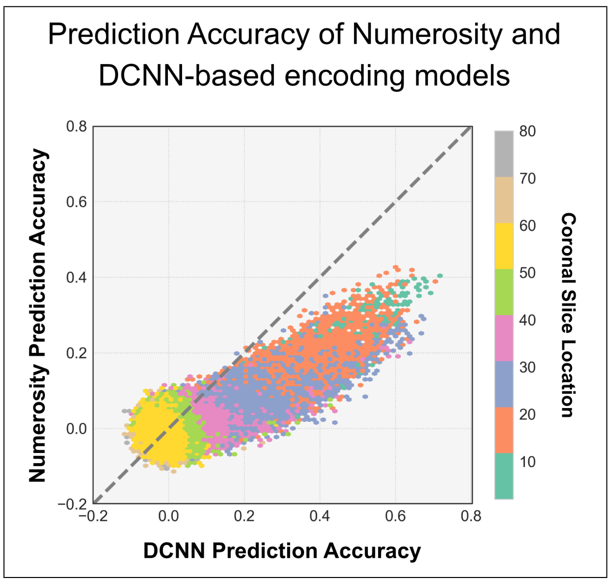

Figure 3: Prediction Accuracy Model Comparison. The joint distribution of prediction accuracy (Pearson correlation between predicted and measured brain activity) for Deep Convolutional Neural Network(DCNN)-based encoding model (x-axis) and Numerosity based encoding model (y-axis). As in Figure 2, voxels are binned into hexagons and colored by average coronal slice location. The numerosity model is completely subsumed by the DCNN model, regardless of slice location, suggesting numerosity may be a feature generated by DCNNs. Data from Subject 1 only.

Conclusion

Our results offer evidence for a representation of numerosity in the human visual cortex during natural scene viewing. This representation appears to be distinct from low-level visual features and maintained primarily in anterior parts of the visual system, in line with previous studies in humans and non-human primates (Harvey et al., 2013; Nieder & Miller, 2004; Piazza et al., 2004; Tudusciuc & Nieder, 2007). Like many other potentially useful and behaviorally relevant representations, the particular representation of numerosity built into our encoding model appears to be subsumed by the features encoded in a performance optimized DCNN. How or if the numerosity representation studied here is utilized to guide behavior or cognition will be an interesting topic for future exploration.

Acknowledgments

NIH R01 EY023384
NSF IIS-1822683

References

DeWind, N. K. (2019). The number sense is an emergent property of a deep convolutional neural network trained for object recognition. *bioRxiv*. doi: 10.1101/609347

Harvey, B. M., Klein, B. P., Petridou, N., & Dumoulin, S. O. (2013). Topographic representation of numerosity in the human parietal cortex. *Science*, *341*(6150), 1123–1126.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., ... Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, *abs/1405.0312*. Retrieved from http://arxiv.org/abs/1405.0312

Nieder, A., & Miller, E. K. (2004). A parieto-frontal network for visual numerical information in the monkey. *Proceedings of the National Academy of Sciences*, *101*(19), 7457–7462.

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*(3), 547–555.

St-Yves, G., & Naselaris, T. (2018). The feature-weighted receptive field: An interpretable encoding model for complex feature spaces. *NeuroImage*, *180*, 188–202.

Tudusciuc, O., & Nieder, A. (2007). Neuronal population coding of continuous and discrete quantity in the primate posterior parietal cortex. *Proceedings of the National Academy of Sciences*, *104*(36), 14513–14518.