A public fMRI dataset of 5000 scenes: a resource for human vision science

Nadine Chang[1] John A. Pyles[2,3] Abhinav Gupta[1] Michael J. Tarr[2,3] Elissa M. Aminoff[4]
{nchang1, jpyles, abhinavg, michaeltarr}@andrew.cmu.edu eaminoff@fordham.edu
[1] Robotics Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213
[2] Department of Psychology, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213
[3] Center for the Neural Basis of Cognition, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213
[4] Department of Psychology, Fordham University, 441 E Fordham Rd, Bronx, NY 10458

Abstract

Vision science - particularly machine vision - is being revolutionized by large-scale datasets. State-of-the-art artificial vision models critically depend on large-scale datasets to achieve high performance. In contrast, although large-scale learning models (e.g., models such as Alexnet) have been applied to human neuroimaging data, the image datasets used on neural studies often rely on significantly fewer images. The small size of these datasets also translates to limited image diversity. Here we dramatically increase the image dataset size deployed in an fMRI study of visual scene processing: over 5,000 discrete image stimuli were presented to each of four participants. We believe this boost in dataset size will better connect the field of computer vision to human neuroscience. To further enhance this connection and increase image overlap with computer vision datasets, we include images from two standard artificial learning datasets in our stimuli: 2,000 images from COCO; 2 images per category from ImageNet (~ 2000). Also included are 1,000 hand-curated scene images from 250 categories. The scale advantage of our dataset and the use of a slow event-related design enables, for the first time, joint computer vision and fMRI analyses that span a significant and diverse region of image space using high-performing models.

Keywords: Big Data; Neural Networks; fMRI; Scenes

Introduction

Recently, artificial vision models have been introduced as potential proxy model of neural representation. The reason for the inclusion of artificial vision models is self-evident when one considers the leaps in machine vision progress in the last few years. Thus, many recent works have leveraged the feed-forward hierarchical structure in neural networks to their advantage. That is, they compare low/mid/high neural reponses in visual processing extracted via neuroimaging with predicted similar level features in a pre-trained network (network already trained on a dataset for a specific task). Neural networks have shown to be more predictive of neural responses in higher layers in the visual hierarchy (Yamins et al., 2014). Additionally, neural networks have also proven to better model human dynamics underlying scene representation (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016) compared to standard models of scene and object perception, GIST descriptors (Oliva & Torralba, 2001) and HMAX models (Riesenhuber & Poggio, 1999; Serre, Wolf, & Poggio, 2005).

With the success in modeling neural data elicited from tasks ranging from scene understanding to object recognition, the incorporation of neural networks as models and analysis tools for biological vision is unavoidable and imperative. Furthermore, increased visual perception understanding alludes that the study of vision science can no longer be isolated into separate spheres of biological and machine vision. We argue that further progression in vision science will require intertwined biological and machine vision approaches. However, one of the biggest obstacles for integrating across the fields of biological and machine vision is data, more specifically a lack of neural data.

The first data consideration is size. The general success in neural networks can be largely attributed to large-scale datasets. High performing neural networks are trained and evaluated on several standard large-scale image datasets. In contrast, although large-scale learning models have been applied to human neuroimaging data, the image datasets used in neural studies often rely on significantly fewer images - typically a few hundred due to time-constrained experimental procedures.

The second data consideration is diversity. The small size of datasets also translates to a limited diversity of images used in neural studies. The images commonly used in neural studies only encompass a small subset of the entire image space. While object recognition has been studied intensively (Khaligh-Razavi & Kriegeskorte, 2014) and in isolation, the typical amount of object categories are not more than 100 categories. However, image datasets used to train and evaluate neural networks encompass a wide range of naturalistic and realistic images with up to thousands of categories. For example, a facial image for neural studies is generally center focused on a face with no noisy background, while a facial image in most artificial vision datasets contains a rich, complicated, and semantically meaningful background with no guarantees of a centered face.

The small scale of neural data and the lack of image feature diversity inherently limit 1) the ability to compare model and measured neural representations and 2) the amount of data that can by modeled by networks.

We address these two data concerns in our newly gathered slow-event related functional magnetic resonance imaging (fMRI) dataset collected from four subjects. To address data size, we dramatically increase the image dataset size

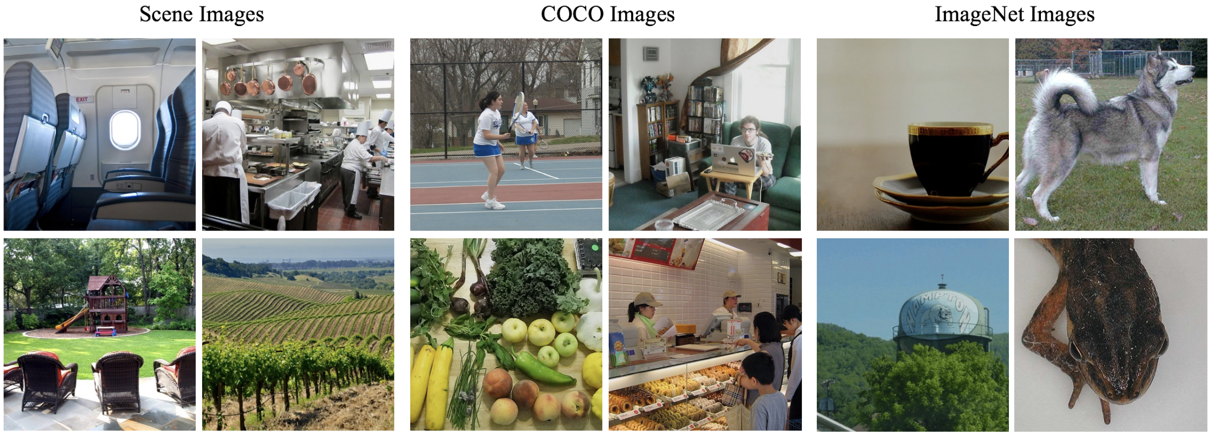

Figure 1: Sample images from each dataset.

deployed in an fMRI study of visual scene processing, scaling the number of images by over an order of magnitude relative to most earlier studies: 5,254 discrete image stimuli were presented to each of four participants. Importantly, the slow-event design allows us to isolate the signal to each individual trial, without any bleed-over from neighboring trials. Thus, our dataset will be widely accessibly without the need for advanced disentangling algorithms. Finally, we will be publicly releasing the dataset to prompt future collaborations between neuroscience and computer vision.

Methods

We will be discussing our methods for data collection in this section.

Stimulus Selection

The visual stimuli presented to each subject is comprised of a total of 5,254 images, of which 4,916 images are unique. The images breakdown into these three datasets: i) 1,000 images from scenes, indoor and outdoor. ii) 2,000 images from the Common Object in Context (COCO) dataset (Lin et al., 2014). iii) 1,916 images from the ImageNet dataset (Deng et al., 2009). Chosen samples used for stimuli from each of the three major datasets are shown in Figure 1.

Firstly, for the scene stimuli, we have 250 unique scene categories chosen mostly from the SUN dataset (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010). We then choose 4 exemplars per category to add to a total of 1,000 scene stimuli.

Second, for the COCO stimuli, we randomly select 2,000 images from COCO training set with a random sampling. The random sample scheme is structured such that it considers the various annotations that accompany each COCO stimulus. Thus, we maintain various image statistics in our sampled data, which ensures that our chosen stimulus set is an accurate representation of the original training set.

Thirdly, for the ImageNet stimuli, we use the standard 1,000 class categories in ImageNet for our image selection. However, due to the extreme affective nature of some image categories, such that they might evoke emotional responses, we remove 42 categories. For each category, we randomly select 2 exemplars per ImageNet category from the ImageNet training set that fulfill our image size and resolution criteria. With 958 categories and 2 exemplars per category, we have a total 1,916 ImageNet stimuli.

All stimuli are RGB and of size 375 x 375.

fMRI Data Presentation

The fMRI data was collected from a total of 4 subjects, (all sessions collected from 3 subjects and half of the sessions collected from 1 subject due to withdrawl from the study). Each subject participated in 16 sessions. All 5,254 images were presented through a total of 15 functional sessions. 4,916 images were presented once, and 113 were presented an additional three times. The remaining session contained anatomical and diffusion scans, and additional localizers.

Each functional session was 1.5 hours long with 9 or 10 image runs. More specifically, there were exactly 8 sessions with 9 image runs and 7 sessions with 10 image runs. In the sessions with only 9 image runs, we included an additional functional scene localizer run at the end of the session. Thus, we had a total of 8 scene functional localizer runs in order to independently define regions of interest and assess data quality and consistency through the study.

During each run 37 stimuli were presented. In order for each runs stimuli to accurate represent the entire image dataset, each runs stimuli dataset category was proportionally the same as the overall dataset. More specifically, in our dataset roughly 1/5th was scene images, 2/5th was COCO images, and 2/5th was ImageNet images. Similarly, the run stimuli break down into 1/5th scenes, 2/5th COCO, and 2/5th ImageNet. Of the 37 stimuli, roughly 2 were repeated images. Thus with 35 unique stimuli per run, 7 were scene images, 14 were COCO images, and 14 were ImageNet images. However, because the total number of images do not divide nicely into 7s, some sessions contained a slightly imbalanced portion of categorical images by a factor of 1 image.

Each run began with a 6 second fixation cross and ended with a 12 second fixation cross. Following the initial fixation cross, all 37 stimuli were shown sequentially. Each stimulus was shown for exactly 1 second followed by a 9 second interstimulus interval.

Each subject was asked to perform a basic valence task for every stimuli. They rated how much they liked each image by making a button response using this metric: 'like', 'neutral', 'dislike'. They responded after the stimuli was presented during the 9 seconds of interstimulus fixation.

fMRI Data Acquisition

Functional MRI data was acquired on a 3T Siemens Verio MR scanner at the Scientific Imaging and Brain Research Center at Carnegie Mellon University using a 32-channel head coil. Functional images were collected using a T2*-weighted echoplanar imaging pulse sequence: 69 slices parallel to the AC/PC; in-plane resolution 2 x 2mm; 2mm slice thickness (no gap); interleaved acquisition; 212mm field of view; 6/8th phase partial Fourier; multi-band factor = 3; TR = 2000ms; TE = 30ms; flip angle = 79 degrees. Each scene run contained 194 volumes, and each functional localizer run contained 141 volumes. To reduce motion and maintain consistent head placement and alignment across sessions, Headcases (CaseForge, Inc.) customized for each participant were used. Data was motion corrected within and across session for further analysis.

Results

We analyze our dataset via two main methods: 1) Representational Similarity Analysis (RSA), 2) Nearest Neighbor. We perform RSA on the correlation between neural responses and AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) features. Neural responses are extracted from regions of interest (ROIs) we have identified through our functional localizers. Nearest neighbor experiments are performed on defined ROIs as well. We perform nearest neighbor within voxel space and visualize results by viewing the stimulus responsible for the neural response.

Discussion

We address one of the biggest obstacles for integrating across the fields of biological and machine vision - data. Thus far, neural datasets are lacking in 1) size, 2) diversity, and 3) stimuli overlap with existing computer vision datasets. We address all concerns in our new dataset where we successfully collect a large-scale, diverse fMRI dataset on 5,254 stimuli that is publicly available. Our data is 1) significantly larger than existing slow-event neural datasets by an order of magnitude, 2) extremely diverse in stimuli, 3) considerably overlapping with existing computer vision datasets.

Additionally, we leverage the magnitude of our data and demonstrate the stability and quality of our data through nearest neighbor. The nearest neighbor results illustrate that we can discern image content form individual scenes. Further, we are able to explore the stimuli relation to other images. The success of our nearest neighbor results is a proof of concept that we have the ability to analyze images through neural data. More importantly, semantically-similar stimuli in top nearest neighbors of various stimuli suggests that we have curated a new set of rich image representations. Similar to how neural networks have been able to provide rich semantically meaningful representations, these neural image representations likewise contain semantics beyond language. Without the restriction and bias of human language, this neural dataset provides the potential to explore visual semantics that have yet to be considered in both neuroscience and computer vision.

Acknowledgments

This work is supported by ONR MURI N000141612007, Sloan, Okawa Fellowship to Abhinav Gupta and also by NSF Award BCS-1439237 to Elissa Aminoff.

References

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*. doi: 10.1038/srep27755

Deng, J. D. J., Dong, W. D. W., Socher, R., Li, L.-J. L. L.-J., Li, K. L. K., & Fei-Fei, L. F.-F. L. (2009). ImageNet: A large-scale hierarchical image database (ppt). *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2–9. doi: 10.1109/CVPR.2009.5206848

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11). doi: 10.1371/journal.pcbi.1003915

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, 1–9. doi: http://dx.doi.org/10.1016/j.protcy.2014.09.007

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 8693 LNCS, pp. 740–755). doi: 10.1007/978-3-319-10602-1$_48$

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175. doi: 10.1023/A:1011139631724

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, *2*(11), 1019–25. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/10526343` doi: 10.1038/14819

Serre, T., Wolf, L., & Poggio, T. (2005). Object recognition with features inspired by visual cortex. In *Proceedings of the ieee computer society conference on computer vision and pattern recognition* (Vol. 2, pp. 994–1000). doi: 10.1109/CVPR.2005.254

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the ieee computer society conference on computer vision and pattern recognition* (pp. 3485–3492). doi: 10.1109/CVPR.2010.5539970

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. doi: 10.1073/pnas.1403112111