CARNEGIE MELLON UNIVERSITY

MASTERS THESIS

Understanding Machine Vision through Biological Vision

Author:
Nadine CHANG

Supervisor:
Abhinav GUPTA

A thesis submitted in fulfillment of the requirements for the degree of Masters of Science in Robotics

in the

Robotics Institute

May 22, 2018

CMU-RI-TR-18-28

CARNEGIE MELLON UNIVERSITY

Abstract

Robotics Institute

Masters of Science in Robotics

Understanding Machine Vision through Biological Vision

by Nadine CHANG

Recent success in machine vision has been largely driven by advanced computer vision methods, most commonly known as deep learning based methods. While we have seen tremendous performance improvements in machine visual tasks, such as object categorization and segmentation, there remain two major issues in deep learning. Firstly, deep networks have been largely unable to adapt to novel yet similar datasets unseen in training time. More specifically, deep neural networks lack robustness. Secondly, there is still a lack of clear understanding at the inner mechanisms that drive deep learning. More specifically, deep neural networks lack interpretability. In this thesis, we explore the robustness of neural networks by generating novel, harder images using Generative Adversarial Networks (GANs). Finally, we propose to address both robustness and interpretability through the incorporation of understandings in biological visual systems. We collect a novel, large-scale functional magnetic resonance imaging (fMRI) dataset in order to gather sufficient data on biological image perception. We show that our dataset contains brain activations that positively correlate to presented visual images.

Acknowledgements

First, I would like to thank my advisor Prof. Abhinav Gupta for his steady, continuous support throughout my Masters study and research. His advice, encouragement, and knowledge have been vital to my growth as a researcher. His steady guidance has allowed me to cultivate my skills such that I am able to produce this thesis.

Second, I would like to thank my collaborators Prof. Michael Tarr, Prof. Elissa Aminoff, and Dr. John Pyles for their guidance throughout my interdisciplinary research. I stepped into this project with no knowledge of Neuroscience. It was only through their continued advice and teachings that I was able to complete this interdisciplinary work.

Finally, I would like to thank my mentor Ishan Misra for teaching me the ropes the in my first year of research. I am also grateful for my labmates Senthil Purushwalkham, Kenny Marino, and Achal Dave for all their support and advice. The copious amounts of research discussions and coffee were no doubt instrumental to my growth.

Contents

Abstract iii

Acknowledgements v

1 Introduction 1

2 Methods 5
- 2.1 Subject Selection and Background . . . 5
- 2.2 Stimulus Background . . . 5
- 2.3 Stimulus Collection . . . 6
- 2.4 Stimulus Presentation . . . 9
- 2.5 fMRI Data Acquisition . . . 10
- 2.6 Nearest Neighbor Analysis . . . 10
  - 2.6.1 Data Format . . . 10
  - 2.6.2 All Stimuli . . . 11
  - 2.6.3 Repeated Stimulus . . . 11
  - 2.6.4 Normalization . . . 12

3 Results 13
- 3.1 Nearest neighbor on all stimuli . . . 13
- 3.2 Nearest neighbor on repeated stimulus . . . 15

4 Discussion 19

List of Figures

2.1 Sample Dataset Images 6
2.2 COCO statistics for number of categories per image 7
2.3 COCO statistics for number of images per category 7
2.4 COCO statistics for number of object instances in per image 8
2.5 ROI Matrix 11

3.1 Subj1 PPA Nearest Neighbor Results 13
3.2 Subj1 Early Visual Region Nearest Neighbor Results 14
3.3 Subj1 Early Visual Region Nearest Neighbor Results 2 15
3.4 Subj1 PPA Nearest Neighbor Repeated Stimulus Results 15
3.5 Subj1 PPA Nearest Neighbor Repeated Stimulus Voxels Results 16
3.6 All Subjects Nearest Neighbor Recalls 17
3.7 Subj1 Normalized Combined Hemisphere Nearest Neighbor Recalls . 18

Chapter 1

Introduction

In the past few decades, vision science has seen tremendous progress in biological and machine vision due to advanced technology. Within biological vision, we have been able to capture complex behavioral neural activity by measuring neural responses in conscious humans. Despite the availability of these neural responses, the mapping of neural activity to visual processing of information in our environment remains an open question. In particular, high-level vision processing and understanding are still difficult to extrapolate from these complex neural activities. Recently, artificial vision models have been introduced as potential better models of neural responses. 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. The advent of big data has propelled significant development in large-scale learning models, particularly deep learning models. These models, most commonly neural networks, are able to achieve high performance in several high-level visual tasks - scene recognition, object recognition, segmentation, detection, and action recognition.

It is undeniable that high-level vision and machine vision perform similar visual tasks. Thus, modeling high-level vision through machine vision is theoretically intuitive. Several works have compared neural data to computer vision models prior to neural networks Leeds et al. (2013). Recently, many previous works have leveraged the feed-forward hierarchical structure in neural networks to their advantage. That is, they compare low/mid/high neural features 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). Previous successful comparisons have been done across human brain activities for object and scene recognition Kriegeskorte (2015); Aminoff et al. (2015). Neural networks have shown to be more predictive of neural responses in higher layers in the visual hierarchy Yamins et al. (2014); Guclu and van Gerven (2015). Additionally, neural networks have also proven to better model human dynamics underlying scene representation Cichy et al. (2016) compared to standard models of scene and object perception, GIST descriptors Oliva and Torralba (2001) and HMAX models Riesenhuber and Poggio (1999); Serre et al. (2005).

With increasing successes in modeling neural data ranging from scene understanding to object recognition, the incorporation of neural networks as models and analysis tools for biological vision is unavoidable and imperative Groen et al. (2018). 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 Tarr and Aminoff (2016). There are two perspectives to consider for data sharing across the two subfields of vision. One, for machine vision, what are the types of neural data that will provide more insight or improvement in machine vision techniques and tasks? Two, for biological vision, what are the types of image data that will provide the best set of neural data that leverages the advantages of neural networks? More specifically, what are the types of neural data needed for the modeling neural activities or for the best comparisons across models and neural representations? Further exploration yields that 3 major data considerations are necessary for successful field integration.

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 and 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.

The third data consideration is image overlap. While many neural representations have been successfully compared to and modeled by neural networks, the types of images used to evoke neural responses are typically divergent from the set of images used to train artificial models. Some images used for neural studies include a complicated and cluttered context, but most of these images are not from

computer vision image datasets. The lack of overlapping images for neural studies and for deep learning restricts the ability to directly compare the neural and model feature representations of standard deep learning evaluative images. In other words, we are unable to explore whether network features or neural encodings are richer or more meaningful in terms of image representation.

We address these 3 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 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. While previous works have gathered large-scale naturalistic fMRI data with movie stimuli Huth et al. (2016); Hasson et al. (2004), no large-scale diverse dataset with isolated slow events currently exist. Thus, all current analysis on large-scale fMRI data requires various techniques to disentangle which stimuli is responsible for which neural response.

To address both data diversity and image space overlaps with computer vision datasets, we include images from two standard artificial learning datasets in our stimuli: 2,000 images from Common Object in Context (COCO) Lin et al. (2014); 2 images per category from ImageNet ($\sim$ 2000) Deng et al. (2009). Also included are 1,000 hand-curated indoor and outdoor scene images from 250 categories largely inspired by a third artificial learning dataset, Scene UNderstanding (SUN) Xiao et al. (2010). SUN, COCO, and ImageNet respectively cover these image domains: indoor and outdoor scenes, objects interacting in complicated realistic scenes, and lastly centered objects in realistic images. These three image collections cover a wide variety of image types, thereby enabling fine-grained exploration into visual representations ranging from natural scenes to human interactions to object categories. Furthermore, all large-scale artificial learning datasets are generally curated by scraping the web for images and later de-noised by humans. In other words, there is a noticeable lack of neural or behavior data on the images that are standard benchmarks for rapidly advancing neural networks. This image overlap across computer vision datasets and neural imaging datasets enables novel neural exploration into these benchmark machine vision images. The size of the neural data on computer vision datasets also allows for novel neural network training directly on neural data, with potential to provide additional insight into network training.

The scale advantage of our diverse 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. While it is clear that a large-scale neural dataset is necessary for integrating across vision subfields, it is imperative to note that a large-scale neural dataset is equally crucial in order to understand how vision is processed and represented in the human brain. On a daily basis, the average human visual system observes a wide range of objects and scenes in complicated backgrounds and perspectives. In order to coherently

and comprehensively encapsulate how vision is processed, the types of images that evoke neural data must reflect the realistic and complex views of the visual domain.

The purpose of this paper is to present this dataset that is publicly available. We present a number of results to demonstrate the strength of this dataset both in reliability and quality, in the application of scene understanding, and in the comparison to neural networks. The results presented in this paper focus on the fMRI blood oxygen level dependent (BOLD) response extracted from specific regions of interest (ROI) that are known to be involved in high level vision. Independent localizers were used to define scene selective ROIs: the parahippocampal place area (PPA), the retrosplenial complex (RSC), and the occipital place area/transverse occipital sulcus (OPA) Epstein and Kanwisher (1998); Park and Chun (2009); Dilks et al. (2013); object selective ROIs: the lateral occipital sulcus Grill-Spector et al. (2001); and a region in early visual cortex. Whole brain data will be available within the accompanying dataset. Our intention of making this dataset available is to provide a resource for both biological vision and computer vision research to use to move the field of vision science forward.

Chapter 2

Methods

We will be discussing our methods for data collection, data stability, and data quality analysis in the following sections.

2.1 Subject Selection and Background

Subjects are recruited primarily from the graduate students in the Psychology Department of Carnegie Mellon University. Due to the long duration of the study, we need to ensure that our subjects are capable of scanning for the full duration of the study with minimum effect on data quality. Thus, we targeted a known population sample that has prior experience with fMRI scanning.

Additionally, subjects report no history of psychiatric of neurological disorders and no current use of any psychoactive medications.

2.2 Stimulus Background

The visual stimuli presented to each subject is comprised of a total of 5,254 images, of which 4,916 images are unique. Images are chosen from three general visual datasets in order to represent higher image diversity. 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. iii) 1,916 images from the ImageNet dataset. Summary is shown in Table 2.1. Chosen samples used for stimuli from each of the three major datasets are shown in Figure 2.1.

We chose these three datasets of images to select from because of the image diversities across these categories. The scenes dataset contains images categories that are inspired and largely taken from the Scene UNderstanding (SUN) dataset meant for scene categorization. In detail, the images are more correlated with scenic like

TABLE 2.1: The number of images per dataset included in stimuli.

Scenes	COCO	ImageNet	Repeated
1000	2000	1916	113

FIGURE 2.1: Sample images from each dataset.

images, with less focus on any particular object, action, or person. We chose scene images from both outdoor (i.e. mountain scenes) and indoor (i.e. restaurant) scenes.

Contrarily, the COCO dataset is geared for a detection task. Due to this complicated task, the dataset contains images that are likewise complicated with several provided annotations. Specifically, images focus on a particular object in a realistic context and frequently its interaction with other inanimate or animate objects. One unique aspect in some of these images is that we are able to observe basic human social interactions. We deem these images as 'social scenes'. The ImageNet dataset is geared for an object categorization task. Thus, ImageNet images tend to have one object as the focus of the picture. Additionally, the object is often centered and clearly distinguishable from the image background.

Summarily, we purposefully compiled a list of images that can be generally characterized as 'scene' images, 'social' or 'complicated' images, and 'object' images. Furthermore, the SUN, COCO, and ImageNet datasets represent some of the most commonly used large-scale datasets for the common visual tasks in the computer vision field. SUN is a standard dataset for scene categorization. COCO is a standard benchmark dataset for object detection. ImageNet is not only a standard benchmark dataset for object classification, but also a popular dataset for pre-training deep learning models, neural networks.

2.3 Stimulus Collection

To ensure the quality of the fMRI data, we must first ensure the quality of the stimuli presented to each subject. Therefore, in the stimulus collection phase, we impose a few filters to select the best possible images for the best possible brain activation. The basic quality checks include image resolution, image size, image blurring, and a hard constraint on RGB images only. Lastly, because sequentially viewing images with different sizes (i.e. a vertical image versus a horizontal image) can cause a change in brain activation, we ensure that all stimuli are square and of equal size. First, we address the method for collecting all 1,000 scenes images. For the scene images,

we have 250 unique scene categories chosen mostly from the SUN dataset. We then need 4 exemplars per category to add to a total of 1,000 scene images. For each scene category, we query Google Search with the scene category name and select from the provided top results. Specifically, after applying filters for a large enough image size and resolution, we filter through the images provided from Google Search. If the image looks clear and free of watermarks, the image is selected. In this manner, we select all 1,000 scene images. The final images are downsized to 375 x 375 pixels, the final size for all stimuli.

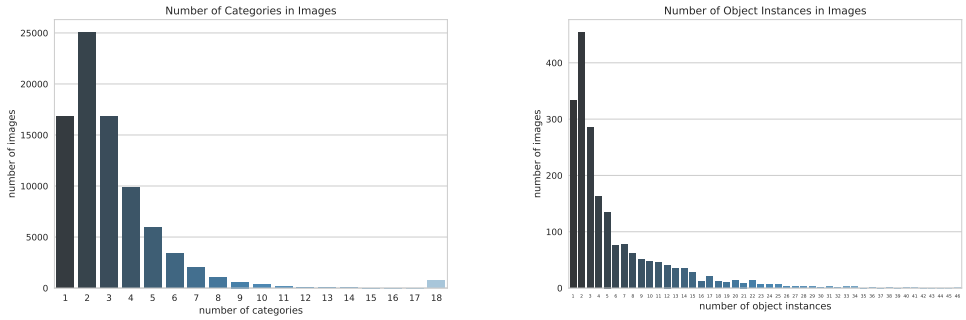

FIGURE 2.2: Here we show the number of images that contain a certain number of categories. Left graph: for all COCO 2014 training set, which we sampled from. Right graph: for all 2,000 images selected from COCO 2014 training set.

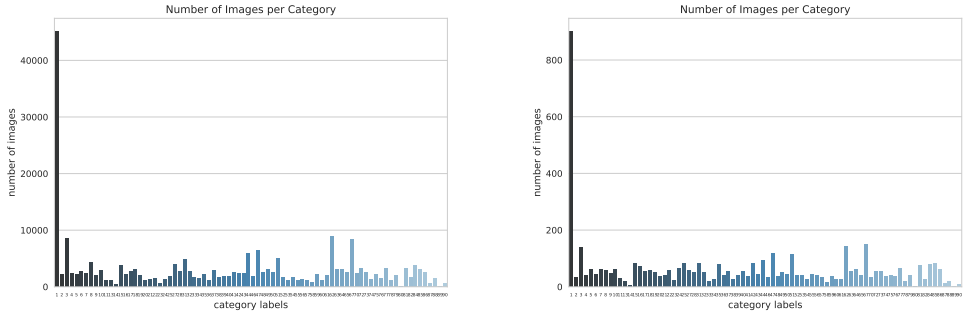

FIGURE 2.3: Here we show the number of images that are in each category. Left graph: for all COCO 2014 training set, which we sampled from. Right graph: for all 2,000 images selected from COCO 2014 training set.

Second, for the COCO images, we randomly select 2,000 images with a random sampling that follows a set of rules. The goal in our random sampling scheme is to ensure that our chosen stimuli set is an accurate representation of the original training set we sample from. Thus, the random sample scheme is structured such that it considers the various annotations that accompany each COCO image. COCO annotations contain 80 object class labels, number of object instances, bounding boxes,

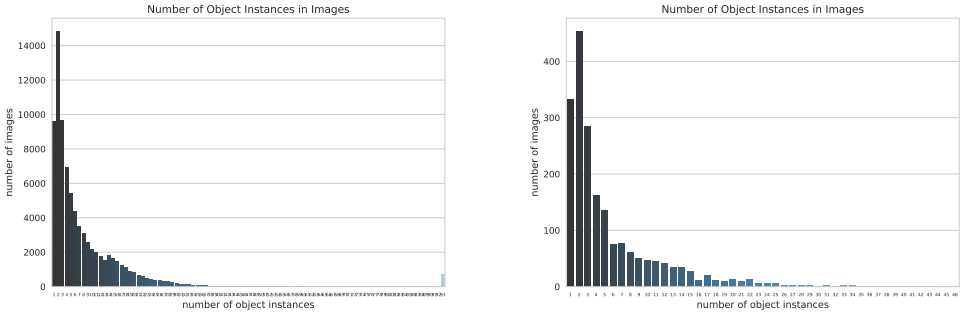

FIGURE 2.4: Here we show the number of object instances in each image. Left graph: for all COCO 2014 training set, which we sampled from. Right graph: for all 2,000 images selected from COCO 2014 training set.

and segmentation polygons. Our final 2,000 images follow these criteria: i) the number of categories in the selected images is proportionally similar to that of the training set as shown in Figure 2.2. ii) the number of images per category is proportionally similar to that of the training set as shown in Figure 2.3. iii) the number of instances per image is proportionally similar to that of the training set as shown in Figure 2.4. iv) the final cropped images contain at least 70% of the original bounding boxes, where boxes are counted if there is an intersection over union of at least 50% between boxes and cropped image. v) the images are bigger than 375 x 375. We go through several rounds of random sampling, where each round we randomly sample according to the above-mentioned criteria before taking a 375 x 375 center crop of the image. However, due to the complicated realistic scenes in these images, often center crops do not contain the main image content. Thus, every passed center cropped image goes through manual inspection. If the center crop contains the relevant image content, the crop is retained. If the center crop does not contain relevant image content, we select a new region of the image to crop from. If there is no reasonable new region to crop from, then the image is rejected. We repeat this process until we reach 2,000 images.

Third, for ImageNet images, we use the standard 1,000 class categories in ImageNet for our image selection. However, due to the violent nature of these images that 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 images. However, ImageNet images all have varying sizes and resolutions. Thus, we only consider images that are bigger than 375 x 375 before taking a 375 x 375 center crop. For all randomly sampled center crops, we manually filter through to determine if a) the crop does not exclude a large portion of the image content, b) the image resolution is high enough. We continue this process until we have exactly 2 exemplars per category.

Lastly, note that we have 5,254 images, with only 4,916 unique images. We randomly select 112 images to be shown exactly 4 times and 1 image to be shown 3 times to each subject. The 113 images are selected such that the image dataset breakdown is proportionally the same as that of the 4,916 images. Specifically, 1/5 of the images are scene images, 2/5 of the images are COCO images, 2/5 of the images are ImageNet images.

Finally, we must consider the RGB and luminance distribution across all of the selected images. Because brain activation is subject to the luminance of the image, we ensure that our images are invariant to this additional factor. To this end, for each image we calculate its hue, saturation, and value (HSV). The value represents the brightness of the image. We find the average brightness per image and find its difference between the gray brightness. We then multiple all values in the image by this new scale. This is otherwise known as grey world normalization. In this way we have ensured the luminance is as uniform as possible throughout all images.

2.4 Stimulus Presentation

The fMRI data is collected from a total of 4 subjects, with only half the data collected for 1 subject and the full data collected for the remaining 3 subjects. Each subject participates in exactly 16 full sessions. All 5,254 images are presented exactly once through a total of 15 sessions. The remaining session contains anatomicals and diffusion scans.

The following session details apply for each subject. Each session is 1.5 hours long with 9 or 10 image runs. More specifically, there are exactly 8 sessions with 9 image runs and 7 sessions with 10 image runs. In the sessions with only 9 image runs, we include an additional localizer run at the end of the session. Thus, we have a total of 8 localizer runs.

The following run and session details apply for each subject. Each run contains 37 stimuli. In order for each run's images to accurate represent the entire image dataset, each run's stimuli dataset category is proportionally the same as the overall dataset. More specifically, in our dataset roughly 1/5th is scene images, 2/5th is COCO images, and 2/5th is 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 are repeated images. Thus with 35 unique stimuli per run, 7 are scene images, 14 are COCO images, and 14 are ImageNet images. However, because the total number of images do not divide nicely into 7s, some sessions contain a slightly imbalanced portion of categorical images by a factor of 1 image.

The following image presentation details apply for each run, each session, and each subject. Before and after each run, in the middle of the black screen seen by the subjects a fixation cross is shown for exactly 6 seconds and 12 seconds respectively. Following the initial fixation cross, all 37 stimuli are shown sequentially. Each stimulus is shown for exactly 1 second followed by a 9 second fixation cross. With

10 seconds spent on each image, we set our repetition time (TR) to 2 seconds. Since each run contains 37 stimuli, we have a total of 370 seconds of stimuli presentation and fixation time. With 6 and 12 seconds of fixation time prior and after the stimuli presentation, we have a total of 388 seconds (6:28 minutes) attributed to each run.

Additionally, each subject is asked to perform a basic valency task for every stimuli. They have to respond how much they like an image using this metric: 'like', 'neutral', 'dislike'. They respond after the stimuli is presented during the 9seconds of interstimulus fixation. They use their fingers to push on a buttons attached through a glove.

Finally, the order of the stimuli presented are randomly selected for each subject. Thus, each subject has a unique stimuli presentation order. The stimuli presentations are also fixed before the start of any sessions for all subjects.

2.5 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 echo-planar imaging pulse sequence. Acquisition spatial parameters: 69 slices parallel to the AC/PC; in-plane resolution 2 x 2mm; 2mm slice thickness; no gap; acquired in an interleaved order; FoV of 212mm; phase partial Fourier scheme of 6/8. Acquisition timing parameters: TR = 2000ms, TE = 30ms, a flip angle of 79 degrees, with 194 measurements for each of the scene scans, and 141 measurements in each of the functional localizer scans. Fat suppression was used. Slices were acquired with a multi-band acceleration factor of 3, no other acceleration factor was implemented.

2.6 Nearest Neighbor Analysis

To demonstrate that our neural data is semantically meaningful in the visual domain, we perform nearest neighbor analysis with euclidean distance on neural data from various regions of interests (ROI). The ROIs we will be examining are PPA, RSC, TOS, LOC, and Early Visual Region. We perform this analysis over all images for all subjects and all TRs. Further, we demonstrate good signal to noise ratio in our data by examining the stimuli corresponding to the 4x repeated images.

2.6.1 Data Format

First, we have to format our data in such a way that it can be properly analyzed. To achieve this, we must also consider all aspects of our neural data that need to be analyzed. More specifically, we need to examine our neural data 1) Intra-subject, 2) Intra-TR, 3) Intra-ROI. Thus, for each subject, each TR, each ROI, we present our data as an N x D matrix, where N is the number of stimuli and D is the number of

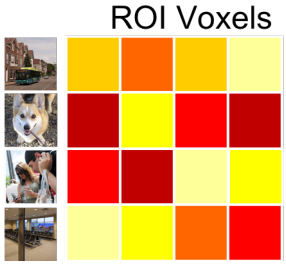

FIGURE 2.5: Sample ROI matrix for nearest neighbors.

voxels of the ROI we are examining. This can be seen in Figure 2.5. In this example, this data represents subject 1 at TR4. We see each row as a vector of voxels for the PPA, and each row corresponds to a unique stimuli.

2.6.2 All Stimuli

First, we look at all ROI data for each stimuli presented to each subject. We explore stimuli to neural data correlation through nearest neighbor on all ROIs neural data of all stimuli. More specifically, for each subject, for each TR, for each of their ROI we perform nearest neighbor analysis on all stimuli present to the subject. In total we perform nearest neighbor on all 5,524 stimuli presented to each subject. We make each stimuli's vector of voxels a query and perform nearest neighbor search over the remaining 5,253 stimuli's vector of voxels. Finally, we visualize the results by viewing the corresponding stimuli for all voxel vectors.

2.6.3 Repeated Stimulus

Second, we explore overall signal to noise ratio in our data by examining the stimulus that are shown 4x to each subject across all 15 sessions. Note that repeated stimulus means that we have 4 unique neural representations for the same stimulus. Since the stimulus is the same, so must be the neural representations, with the exclusion of noise and session to session variance. Thus, we leverage the extra neural representations to our advantage. We do so by examining the nearest neighbor results over the repeated stimulus, which with 4 repeats of 113 images comes to a total of 451 images. We use the 451 images responses as queries and also as the pool of representations to search for its immediate nearest neighbor. Again, we visualize the results by viewing the corresponding stimulus for all nearest neighbors. Here, we examine whether or not one of the repeats of the query image appears in the top k nearest neighbor. We do so by performing basic recalls on all nearest neighbor results. More specifically, if the same stimulus appears in the top k nearest neighbor of a stimuli, then we consider that a positive recall. Thus, for all 451 images, we get a recall of x/451.

2.6.4 Normalization

Finally, we perform all of our analysis described thus far on normalized neural data. We normalize our data by z-scoring across voxels for each stimulus. Let's consider our N x D matrix of voxels. We perform z-score normalization column wise on each N x D matrix.

Chapter 3

Results

We will be discussing data quality analysis in the following sections. We show that our neural data contains rich semantic information about the corresponding stimuli. We demonstrate this through nearest neighbor evaluations, both qualitatively and quantitatively.

FIGURE 3.1: Sample nearest neighbor results for subject 1's PPA at TR 4.

3.1 Nearest neighbor on all stimuli

Through qualitative results, we show our results on nearest neighbor performed over all 5,254 stimuli intra-subject, intra-TR, intra-ROI. Through our results, we see that the top 10 nearest neighbors contain semantically similar stimuli to that of the query stimuli. For example in Figure 3.1, we examine stimulus to voxel correlation through the PPA. We see the monkey stimulus query has other animal stimulus appear in its top 3 nearest neighbor. Additionally, we observe human sport stimuli stimulus has other human sport stimuli stimuli in its top 3 nearest neighbor. It is

imperative to note that while the same sport stimulus does not appear in its top nearest neighbor, the fact that other sport stimulus appear implies that general semantics is captured more significantly than fine-grained semantics. Finally, we observe that cluttered indoor seating stimulus query has similar indoor scenes in its nearest neighbor.

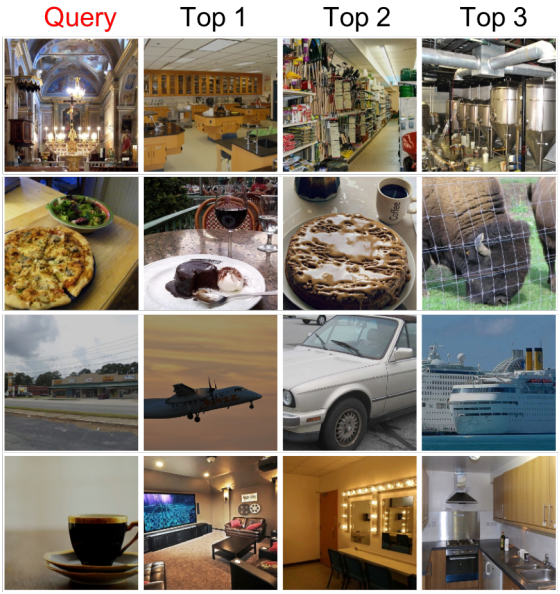

FIGURE 3.2: Sample nearest neighbor results for subject 1's Early Visual Region at TR 4.

Similarly, we observe results from the Early Visual Region. Although, the Early Visual Region is not known for high level semantic processing, we observe semantically similar stimuli to that of the query stimulus. For example in Figure 3.2, we see food stimuli query have other food stimuli in its nearest neighbor. Additionally, we observe visually similar stimuli to that of the query stimulus. For instance, we observe indoor scene stimulus query with other indoor scene stimuli in its nearest neighbor.

However, as demonstrated in Figure 3.3, we still observe noise in our neural data. We observe several nearest neighbor results with seemingly random visual correspondence. This suggests that additional noise pruning may be necessary to increase our signal to noise ratio.

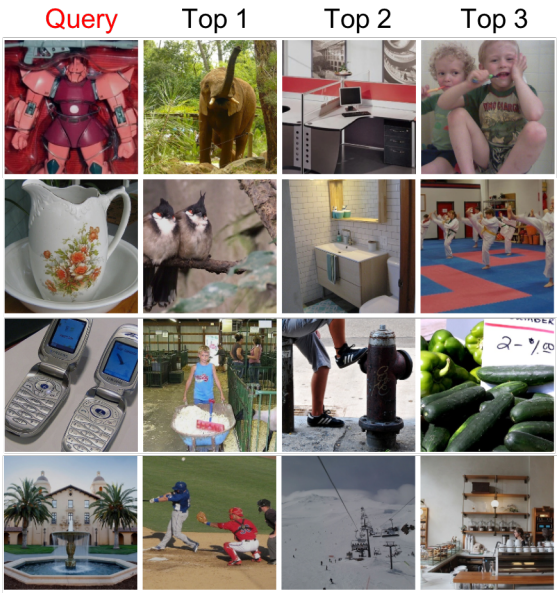

FIGURE 3.3: Sample noisy nearest neighbor results for subject 1's Early Visual Region at TR 4.

3.2 Nearest neighbor on repeated stimulus

Here, we show our qualitatively and quantitative nearest neighbor results on averaged repeated stimulus' voxel representations.

FIGURE 3.4: Sample nearest neighbor results for repeated stimulus for subject 1's PPA at TR 4. We see the same image in the query's immediate nearest neighbor.

In our results, we demonstrate that signal is stable across sessions. We show this by observing for a few stimuli that the same stimulus appear in the repeated stimulus's nearest neighbor as seen in Figure 3.4. This indicates that the noise across sessions does not affect the overall neural representation of an image. Although some stimuli queries do not contain its own stimuli in its top nearest neighbor, we observe semantically similar stimuli in those positions as seen in Figure 3.5.

FIGURE 3.5: Sample nearest neighbor results on repeated images ROI voxels for subject 1's PPA at TR 4.

Finally, we show quantitative results on recalls for the top 10 nearest neighbor. We calculate our recalls based on whether or not the stimulus appears in its own top 10 nearest neighbor. In Figure 3.6, we see high recalls for all subjects.

Note that chance is

$$1 - \binom{447}{10} / \binom{450}{10} = 0.653$$

Furthermore, we observe a clear increase in recall as we approach TRs 2 and 3. In the case of subject 1, we see highest recalls at TRs 4. Additionally, we show that combining hemispheres also leads to an increase in recalls in Figure 3.6. However, note that while z-scoring the voxels generally leads to higher recalls, we see more performance boosts in earlier TRs in Figure 3.7. Thus, it is unclear whether normalizing assists in boosting signal to noise ratio.

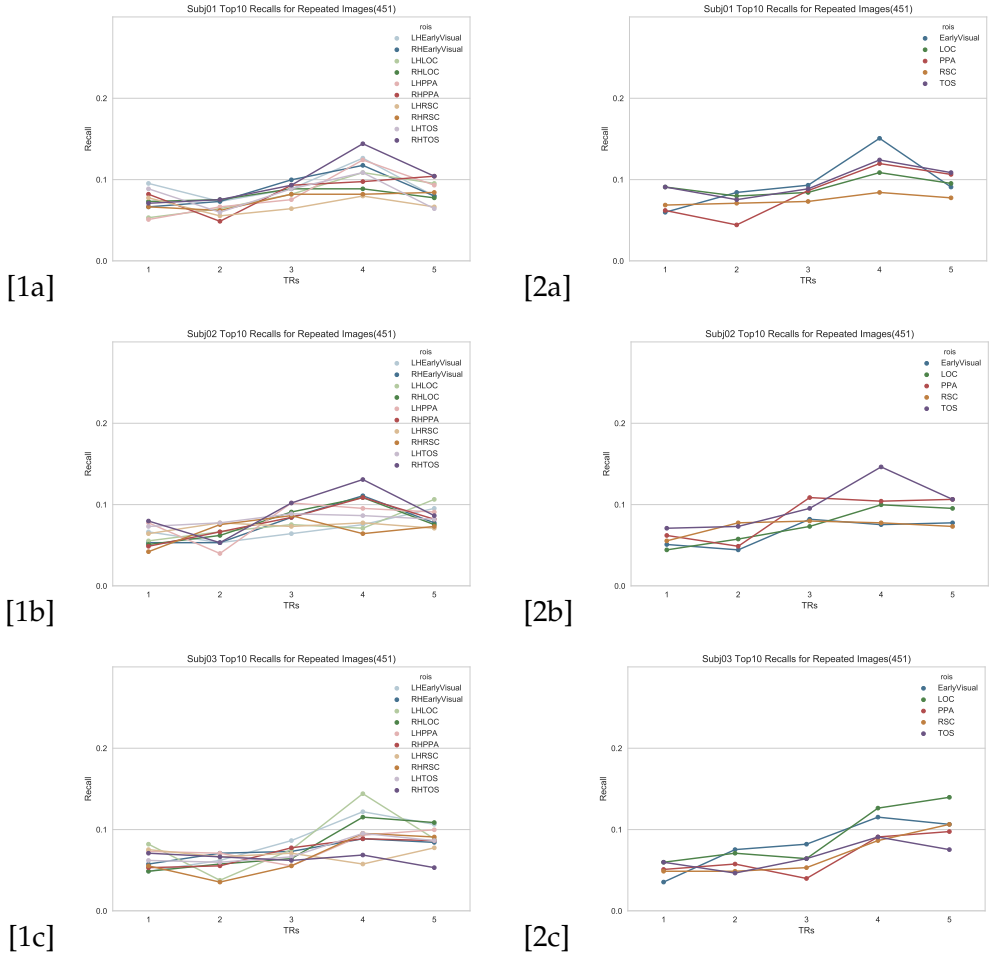

FIGURE 3.6: Recalls for nearest neighbor on repeated images ROI voxels for all subjects. Column 1 shows results for separated hemispheres. Column 2 shows results for combined hemispheres.

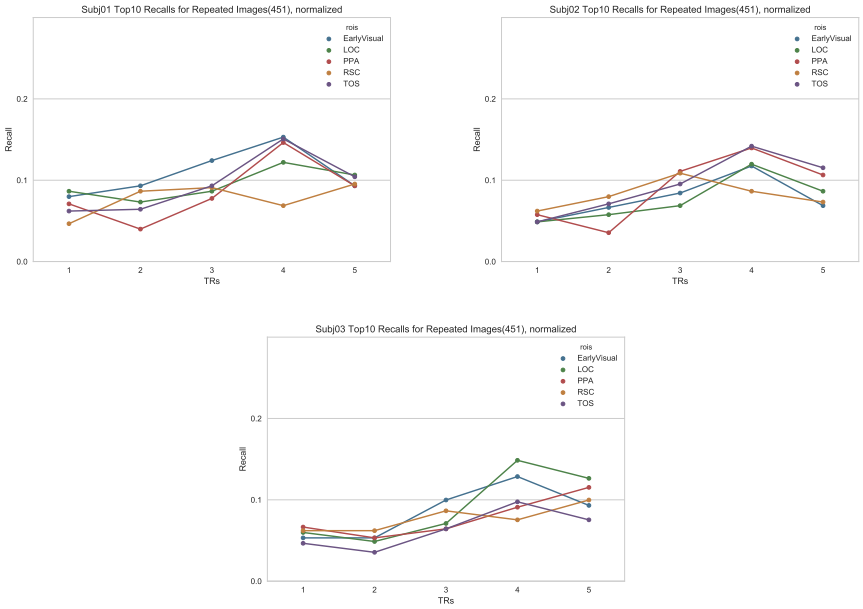

FIGURE 3.7: Recalls for nearest neighbor on repeated images' ROI voxels for combined hemispheres. Voxels are normalized before nearest neighbor analysis.

Chapter 4

Discussion

Further progression in vision science will require intertwined biological and machine vision approaches. In this paper, 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.

Bibliography

Aminoff, E. M., M. Toneva, A. Shrivastava, X. Chen, I. Misra, A. Gupta, and M. J. Tarr
2015. Applying artificial vision models to human scene understanding. *Frontiers in Computational Neuroscience*, 9.

Cichy, R. M., A. Khosla, D. Pantazis, A. Torralba, and A. Oliva
2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6.

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

Dilks, D. D., J. B. Julian, A. M. Paunov, and N. Kanwisher
2013. The Occipital Place Area Is Causally and Selectively Involved in Scene Perception. *Journal of Neuroscience*, 33(4):1331–1336.

Epstein, R. and N. Kanwisher
1998. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601.

Grill-Spector, K., Z. Kourtzi, and N. Kanwisher
2001. The lateral occipital complex and its role in object recognition. In *Vision Research*, volume 41, Pp. 1409–1422.

Groen, I. I., M. R. Greene, C. Baldassano, L. Fei-Fei, D. M. Beck, and C. I. Baker
2018. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife*, 7:e32962.

Guclu, U. and M. A. J. van Gerven
2015. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27):10005–10014.

Hasson, U., Y. Nir, I. Levy, G. Fuhrmann, and R. Malach
2004. Intersubject Synchronization of Cortical Activity during Natural Vision. *Science*, 303(5664):1634–1640.

Huth, A. G., W. A. De Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant
2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

Khaligh-Razavi, S. M. and N. Kriegeskorte
2014. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11).

Kriegeskorte, N.
2015. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 1(1):417–446.

Krizhevsky, A., I. Sutskever, and G. E. Hinton
2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, Pp. 1–9.

Leeds, D. D., D. A. Seibert, J. A. Pyles, and M. J. Tarr
2013. Comparing visual representations across human fMRI and computational vision. *Journal of Vision*, 13(13):25–25.

Lin, T. Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick
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)*, volume 8693 LNCS, Pp. 740–755.

Nishimoto, S., A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant
2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646.

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

Park, S. and M. M. Chun
2009. Different roles of the parahippocampal place area (PPA) and retrosplenial cortex (RSC) in panoramic scene perception. *NeuroImage*, 47(4):1747–1756.

Riesenhuber, M. and T. Poggio
1999. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–25.

Serre, T., L. Wolf, and T. Poggio
2005. Object recognition with features inspired by visual cortex. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, Pp. 994–1000.

Tarr, M. J. and E. M. Aminoff

2016. *Big Data in Cognitive Science*, chapter Can Big Data Help Us Understand Human Vision?, Pp. 343–363. Psychology Press.

Xiao, J., J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba

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.

Yamins, D. L. K. and J. J. DiCarlo

2016. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

Yamins, D. L. K., H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo

2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.