Investigating and Identifying Food Selective Regions in the Brain [DRAFT]

Nidhi Jain

CMU-CS-22-108

May 11

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Leila Wehbe, Chair
Michael J. Tarr

Submitted in partial fulfillment of the requirements for the degree of Masters of Science.

Abstract

Existing research has demonstrated functional selectivity in the brain for high level categories of faces, places, bodies, and sensory and motor processes. Food, despite its abundance and importance, has not been considered as a category for which there is a visual selective region, perhaps because of its lack of visual coherence. In this paper, we investigate responsiveness to food in a large scale natural setting via several statistical methods performed on high-resolution fMRI response dataset to natural scenes. We identify two regions consistent across all participants in the high level visual cortex that appear to be functionally selective for food.

Acknowledgments

TODO

Contents

List of Figures

Chapter 1

Introduction

The human brain is arguably the most complex information processing system to exist, and understanding it has remained a Herculean yet impactful feat. Emerging technology in industry has hoped to better understand the brain in efforts to create representative simulations and interpretable brain-machine interfaces [1, 2]. Biological fields are motivated to understand the brain in order to directly help patients with neurological disorders. Deep learning researchers benefit from findings about the brain to inspire novel model architectures, similar to how key components of Reinforcement Learning and Computer Vision have been inspired by discoveries of certain brain function characteristics [3, 4]. The impacts of better understanding the brain are clearly ample.

One line of significant research has furthered the understanding of the brain through investigation of its structural organization; this research has shown promising results. Studies have discovered the existence of various regions that tend to selectively respond strongly to specific categories. One notable region for example, the Fusiform Face Area (FFA), has been found to often respond selectively to the human face [5]. In addition to the FFA, research has discovered the places, bodies, and visual words also correspond to selective regions in the brain despite their high-level cognitive characteristics [6].

Little research has considered the category of food as a potentially visually selective area past discovering certain areas activated by food. Food remains a fascinating category due to its lack of visual coherence across the category. In other words, there are little to no visual characteristics that remain uniform across the category of food, yet we are still able to recognize unseen foods as edible. This characteristic highlights the importance of the high-level processing to this potential region, and shines light on the fascinating aspect of a possible selectivity despite such a diverse category. Selectivity for this category goes beyond just food-related impact, and would emphasize that the brain has the ability to have a functional unit that is able to combine visual, semantic, and top-down influences to unify a diverse set of items.

In this work, we identify and investigate an apparent functional selectivity for the food category in the brain. We do so by performing several machine learning and statistical methods on brain activation data to uncover consistent patterns that point to a specific cortical region's

functional selectivity. These techniques include encoding models that predict participant brain activity given the images that participants viewed, decoding models that predict viewed images' visual characteristics based on participant brain activity, and more. We use statistical methods to further break down the structure and functionality of this proposed food region. We then use state of the art deep learning models to confirm that patterns found in brain activity responses to food are not solely driven by low level visual features of the input images that participants viewed.

The identification of a food selective region has the potential to help inform how one would detect food in an environment, as well as how to organize food in a semantic system. It helps provide further insight into understanding how the brain may represent life-sustaining objects, and what kind of system we use to process food (whether it is ancient or not). Among many, some impacts on modern technology include improving food classifiers with the addition of brain response data, helping build reinforcement learning agents that seek out food, and identifying whether food identification could be a shared learning task vs a specialized task for robotics applications.

Chapter 2

Background

The representation of high-level visual information in the human brain has been marked by the phenomenon of selectivity for categories of high ecological importance. There are multiple brain regions that show preferential neural responses to such visual categories: faces, bodies, and places [7, 8, 9, 10]. Independent of any particular theory for *how* these functional brain regions arise [11, 12], we can agree that these regions exist because they instantiate processes and representations for categories that are highly relevant for day-to-day behavior. In a similar vein, food is a category that we might expect to be evolutionarily relevant – finding nourishment being more ancient than social interaction and, arguably, more fundamental to survival. Thus it is somewhat surprising that food has not been identified as a visual category for which localized preferential neural responses are observed.

While it is known that the visual presentation of food images prompts a range of neural responses, including both cognitive and emotional effects, the concurrence between studies using food images has been quite low [13]. For example, while some food-related responses have been observed bilaterally in the fusiform gyrus and left laterally in orbitalfrontal cortex, only 41% of the studies included in a meta analysis contributed to the significant food/non-food contrasts [13]. In addition, the majority of studies comparing food to non-food have focused on the interaction between food processing and physiological factors such as body weight or hunger state. As such, although food images interact with both food selection and regulation of food intake [13, 14], there have been no clear tests as to whether food elicits preferential responses in the human visual system.

An important factor that may influence prior studies of food-elicited neural responses is context. One of the claims regarding category-selective neural responses is that they are elicited automatically by the category in question [**?**]. In contrast, in many of the studies using food images the manipulations used appear to be sensitive to task demands (e.g., affective state influences attention and concurrent early visual responses to food [**?**]). We posit that by focusing on isolated food and non-food "posed" visual stimuli, many studies have failed to consider food in environmentally relevant contexts. As such, decontextualized food images – which vary widely in visual appearance – may not be sufficiently salient or realistic enough so as to elicit food-related processing by default (although such images may do so with task-driven attention). In

contrast, both faces and bodies exhibit very little visual variation within category, while places are typically tested using natural scene images. In a nutshell, the extreme within-category variability in food appearance may contribute in several ways that may render identifying food-preferential brain regions more challenging than other ecologically important categories. First, without context, *images* of food may not present as food *qua* food to the visual system. That is, category membership for images of food may be ambiguous absent the presence of associated semantic information that helps, possibly in a top-down manner, constrain and narrow the categorical content of the image. Second, because of the high visual variability for food as a category, detecting significant food-driven responses may require more sensitive designs than provided by standard neuroimaging designs (which typically rely on small numbers of images and trials per condition and therefore depend on high within-condition similarity).

Our present study addresses these issues in two significant ways. First, real-world images, drawn from the the COCO dataset [15], were used for both the food and non-food conditions. Second, functional MRI (fMRI) data in response to viewing these images was collected on a massive scale [16], thereby improving our ability to detect effects across conditions. To preview our most important result, we reliably identify two distinct regions in high-level visual cortex that are preferentially responsive to food images. As with other robust category preferences, we take our results to indicate that the human visual system instantiates (through unknown mechanisms) processes and representations that support perceiving and reasoning about food. In that food is incontrovertibly an ecologically critical category, we view this finding as highly consistent with earlier findings of preferential mechanisms for the perception of faces, bodies, and places.

Chapter 3

Methods

3.1 Dataset

In order to investigate responsiveness to food in a large scale natural setting, we take advantage of the Natural Scenes Dataset (NSD), a dataset of high-resolution fMRI responses to natural scenes via a recognition task [16]. These natural scene images are pulled from the annotated Microsoft Common Objects in Context (COCO) dataset [15]. NSD has brain response data from 8 screened subjects that each see roughly 9500 natural scene images over the course of a year. These viewings were administered during 30-40 scan sessions. Of the 70,566 total images viewed across participants, 1000 of these images are viewed by all 8 participants. We took special interest in these 1000 images' response data when conducting analysis in order to ensure consistency in results.

3.2 Labeling

As outlined in Figure 1, we use a categorical hierarchical structure to hand label the 1000 shared images viewed by all subjects with their location classification, object classification (including the binary existence of food), and perspective classification of the image. Image perspective is discretized into 'zoom,' 'reach,' or 'large-scale.' 'zoom' signifies an apparent zoom in of the camera lens, thereby likely focusing on one object and excluding surrounding information. 'reach' images demonstrate affordances by displaying objects at a human-reachable distance [17]. 'large-scale' images encompass the remaining images, which likely include an image of a general scene as opposed to one or more close up objects. The image perspective category's vague nature leaves it vulnerable to variation in labeling. To avoid this variation and ensure consistency, we underwent several rounds of labeling and verification. We also ensured that these shared images followed similar labeling distributions to the non-shared images, as shown in Figure 0.

Figure 3.1: The 1000 images viewed by all 8 subjects in NSD were manually relabeled in order to investigate responsiveness to natural food images. The top half of this figure shows example images that were labeled as well as their corresponding labels, while the bottom half demonstrates the organizational structure used when labeling a given image. Each of the images were given at least one label within each of the three categories of location, content, and image perspective. Multiple content labels could be and often were used for a given image

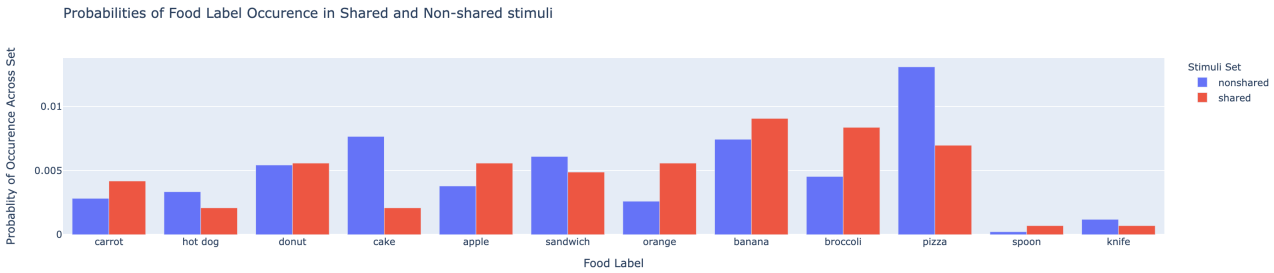

Figure 3.2: The images viewed by all subjects (shared) were expected to have a similar label distribution overall to the remaining images (nonshared). However, we confirm that the food-label distribution specifically remained similar across both sets. We plot the shared (red) and non shared (blue) percentage of images corresponding to each label. This distribution similarity allows us to ensure that findings are not due to data distributions specific to a set.

3.3 Encoding and Decoding models

To identify voxels especially responsive to certain labels, we encode all 16 labels into a single binary vector per image and perform voxel-wise ordinary least squares encoding models. We then run statistical significance tests between pairs of weights, where each weight is the model coefficient for a different label input.

While an encoding model is able to provide some insight into single-voxel selectivity through response predictions, a decoding model can uncover distributed pattern-level representations of visual features. As a result, we perform a voxel-wise searchlight algorithm to classify the existence of food and further inspect high-performing regions.

3.4 Understanding Fine Grained Region Structure

Encoding and decoding models help identify proposed regions especially responsive to food, but are unable to provide further insight regarding the structure of these 'food regions'. We run a principal component analysis to better understand the structure and correspondence in these food regions. To verify our proposed food regions resulting from analysis on the 1000 shared images, we perform further investigation across the remaining images that were not already used. These remaining images are no longer hand labeled, so we use we use a ridge regression model on COCO labels and visualized the resulting voxel-wise weights per label.

In addition to investigating the weights of the ridge regression, we directly consider the non-shared images' activation patterns in our proposed food region. Using the proposed region as a voxel mask, we extract each image's corresponding food-voxel activity. We then perform a K-means clustering algorithm on this activity and investigate the resulting groups. In order to better understand potential visual and semantic justifications behind these voxel-based clusters and isolate these justifications from top-down cortical influences, we also cluster visual and semantic embeddings of these images and look for possible correlations between the image clustering results and voxel clustering results. To obtain these visual and semantic embeddings we use two

state of the art models, CLIP and ResNet-50 [18, 19]. CLIP, trained on both images and text, allows us to extract features with semantic and visual meaning. ResNet-50, training on solely images, results in purely visual feature-based embeddings.

Chapter 4

Results

4.1 Encoding Models

We find two main areas in the visual cortex that are consistently selective across all subjects for food when compared to other non-food categories, as shown in the top half of Figure 2. This finding persists across all 8 subjects. Running significance tests comparing various pairs of labels' corresponding encoding model weights leads to the recognition of 2 regions highly activated by food in the visual cortex. When considering a one sided significance test between 2 labels, for example, food and face, the significant voxels that achieve a p-value lower than 0.05 are identified as more responsive to food than faces. As in Figure 2, the food-significant voxels are often extremely close to 0, and thus highly significant, when compared against other categories. These voxels thus seem to be most activated by the existence of food in the stimuli as opposed to other categories. It is important to note that this spatial 'food' pattern persists even when compared against reach spaces.

As a sanity check, we verify that performing a similar significance test for faces reveals the fusiform face area (FFA) [8]. We choose the FFA here due to the high occurrence of faces in the dataset and reliability of this region. It is clear that these proposed 'food-significant' voxels are consistently more reactive to food stimuli than other stimuli .

4.2 Decoding Models

The high-accuracy voxels from the searchlight decoding model overall align spatially with the significant voxels from the encoding model, as is clear in Figure 2. We note that high-accuracy voxels do not necessarily correspond to food-selective regions, but rather to any region which helps identify whether or not the given stimuli contains food. So, high-accuracy voxels can signify both a food-selective region and an anti-food selective region that especially helps determine that the stimuli is not food-related. Running a voxel-wise searchlight classification allows an investigation into food-reactive voxels from a different perspective from the previously mentioned encoding model. The inclusion of the previously identified food regions in the high-accuracy searchlight voxels further confirms our proposed food regions. Similar regions as those found

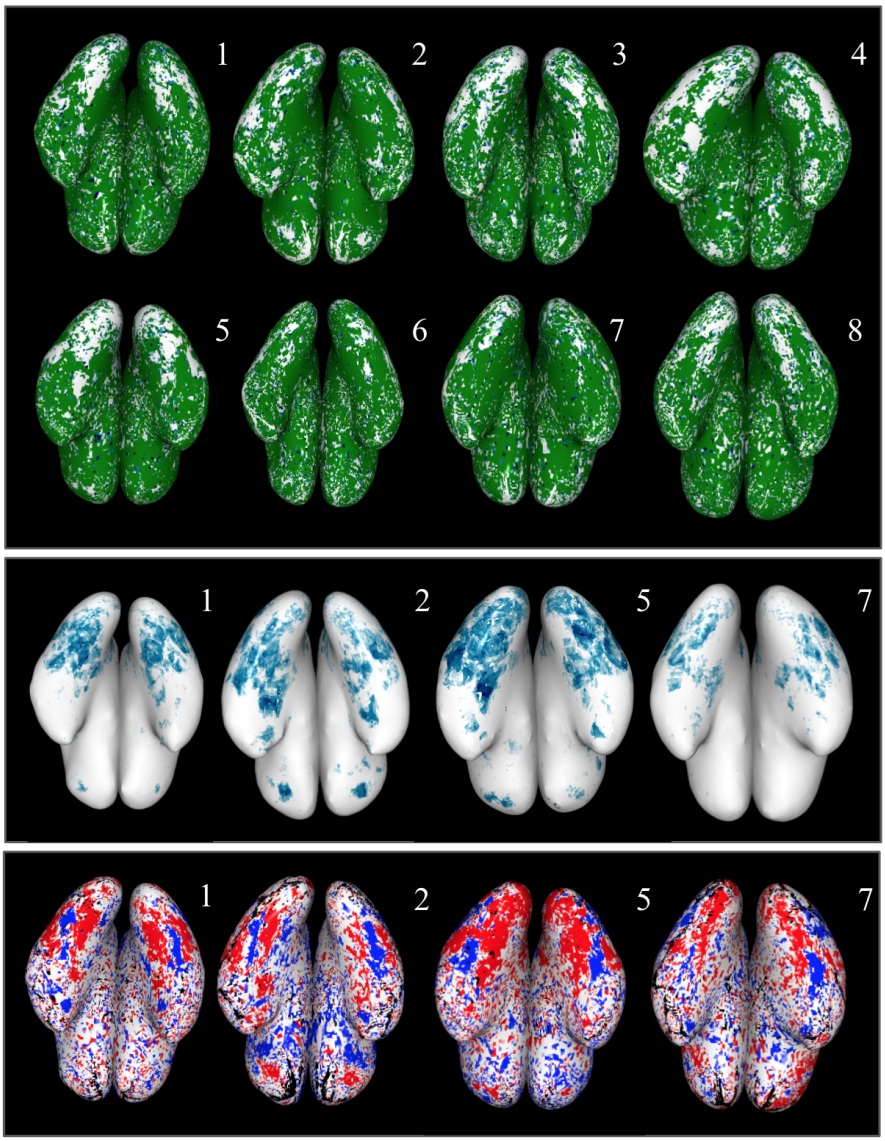

Figure 4.1: This figure shows voxels that are determined to be food-related based off of two differing methods. The top image shows significant voxels (white) from a 1-sided t-test comparing food weights from a trained OLS model against all other non-food object weights. The middle image shows classification accuracy for voxel-wise searchlight decoding, with darker blue voxels signifying higher accuracy. The bottom image shows significant voxels from a 1-sided t-test comparing food weights against face weights (red) as well as the results for a 1-sided t test comparing face weights against food weights (blue). It is notable here to compare significant 'food' voxels to 'face' voxels because the of the high number of face images in our dataset, as well as the robust reliability of the FFA. The food-related regions resulting from both methods overall align with each other, suggesting that these regions are reliably more responsive to food stimuli than other categories.

in the encoding model are observed to be highly correlated with the stimuli food classification accuracy.

Both the encoding and decoding models confirm that these regions seem to be consistently more reactive to food than other categories. However, these models provide little insight into the more fine grained semantic and spatial structure of the responses within this food region. To discover these intra-region patterns, we run a principled component analysis (PCA) on the selected visual cortex regions and investigate the relevant contributing images.

4.3 Semantic Regions Axes

Running PCA on the hypothesized food region leads to a further spatial breakdown of various 'food' areas, as is shown in Figure 3. These breakdowns are able to account for a significant portion of the variance of this region's responses. Each of the uncovered components' axes seems to correspond with a general semantic pattern of image content. To achieve these components, we perform PCA on the matrix of the proposed food regions' voxels by the 108 food-related images from our dataset. The top components explain 34.31, 12.68, and 11.16 percent of the variability, respectively. These components, each appearing to correspond to varying semantic characteristics, have unique contributions to this region's activity. These results demonstrate that explaining the variance of this proposed food region can be done via components with unique correspondences to semantic categories.

4.4 Generalizing Across all Stimuli

We further consider voxel responses to food by inspecting how voxels react to specific food categories from images not yet inspected when proposing the food region. In order to use images not yet inspected, we consider the remaining images that were not viewed by all 8 subjects. We run a voxel-wise ridge regression on the provided COCO labels on these images to predict voxel activity. Then, we visualize the resulting weights per label on each voxel, as shown in Figure 4. In order to observe the which voxels' prediction accuracies benefit most from the addition of food labels, we perform the ridge regression on all labels, as well as all but food labels. We then compare the R^2 values voxel-wise. We observe one general activation pattern by the food-related labels. This pattern corresponds strongly with our previously proposed food region, and aligns with the subset of voxels whos' prediction performance improves due to the addition of food labels. This improvement suggests that these voxels are more responsive to food than other cortical areas. Activations across the food labels are extremely consistent. The responses re-emphasize our identification of a food-selective region.

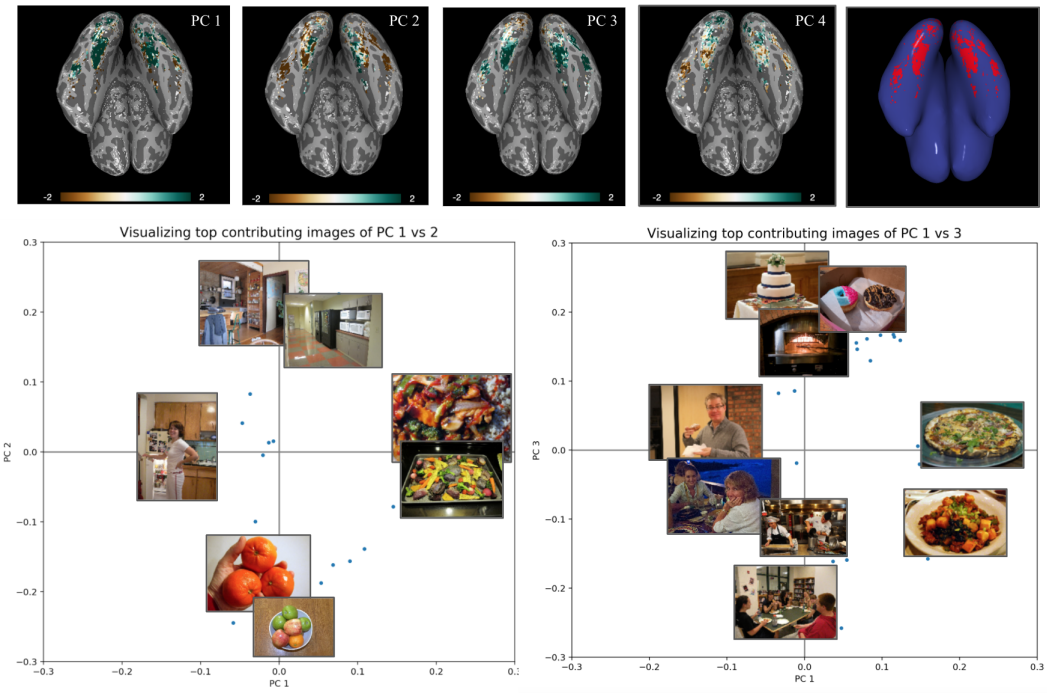

Figure 4.2: Applying PCA on the proposed food region's responses provides more insight into the structure of these responses. Above, we can see the different regions corresponding to different components, as well as the top images for each component. The top images display each components projection onto the brain, demonstrating the generally positively (green) and negatively (brown) contributing voxels. On the top right, we see in red the region that we have identified as the relevant food region, generated based off of the intersection of the food vs face encoding model test and manually identified visual cortex regions. In the graphs, we see stimuli that most contribute to various components, as well as their placement. We see that each component overall highlights different semantic characteristics of the visual stimuli. The first component appears captures the surface area of food in the image, the second distinguishes places, and the third seems to capture a combination of spatial frequency and people. These semantic categories are able to explain a large component of the variance resulting from these food regions.

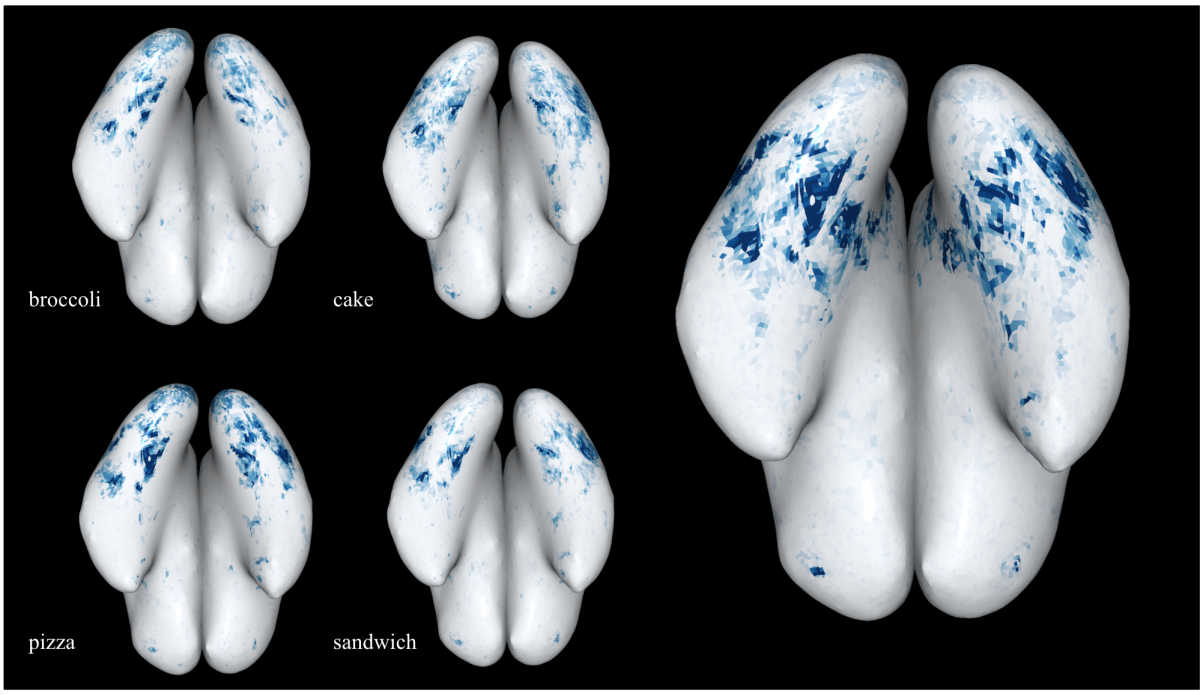

Figure 4.3: Analysis on predicting voxel activity from unseen images points to the same food-selective regions as previously proposed. We observe 4 COCO label weights corresponding to each voxel following a voxel-wise ridge regression on COCO labels (left) We also observe the voxel-wise improved R^2 values from including food labels when predicting voxel activity (right). Both the weights corresponding to individual food labels as well as the overall impact of the aggregate food labels highlight our proposed food regions. These results also point to the conclusion that this region is food selective.

Figure 4.4: We observe the highest voxel-activity inducing images in each cluster. This helps us identify how our proposed food region reacts to different images. On the left we see in red the voxels identified as the proposed food region, and thus used here to cluster the images. In the middle we see top images corresponding to our first cluster, and on the right we see the top images corresponding to the second cluster. We clearly this region reacts differently to food than to other images.

4.5 Voxel Embedding Clusters vs Image Feature Embedding Clusters

To confirm that our proposed region does in fact have this unique food selection, we test whether our proposed food region responds to new food-related images significantly differently from new non-food related images. In this test, our new subset of images also includes an even split of non-food and food images from images that were not used to determine the originally proposed food region. Using the previously proposed mask of voxels identified as food-selective, we create a subset of corresponding voxel activity for each image. We then use the K-means algorithm to cluster the voxel data corresponding to these images to reveal patterns of reactivity in this region. 2 clusters emerge, each containing voxel activity corresponding to an image. We now consider the images corresponding to this voxel activity, and thus the 2 clusters of images. We notice that 1011 of the 1638 food images gather in one cluster, and observe the images with the highest voxel activity in each cluster (Figure 5). The top images in each cluster have significant semantic differences, with one cluster clearly corresponding highly with food. These top cluster images further confirm our proposal of this region as selective to food.

We use voxel responses to further cluster food images and observe the resulting clusters. Especially when considering the visual cortex, it is important to isolate causes for variations in brain responses from solely image feature variation. To help understand if variations in food responses are due to visual characteristics, we use CLIP and ResNet to cluster food-related images and observe resulting patterns. We then observe whether these patterns are consistent with the clusters emerging from voxel-based image clustering. CLIP considers both visual and semantic features when training, while ResNet only considers visual features. In both outputs, we notice semantic clusters that do not have clear correspondence with our resulting food clusters (Figure

6), suggesting likely top down processing that goes further than raw visual features.

Figure 4.5: We observe image clusters based on voxel-response embeddings (top). We then compare this to image clusters based on CLIP embeddings (bottom). Two clusters emerge from voxel-response embeddings, while 4 main clusters emerge from CLIP. The two clustering patterns show little to no correlation with each other, demonstrating that this proposed food region involves more complex, top down processing rather than solely low level or semantic features.

Chapter 5

Conclusion and Future Work

How are knowledge representations organized in the human brain? Within the visual system, one of the hallmark discoveries of the past several decades has been *category selectivity* [7, 8, 9, 10] We saw here results that clearly suggest selectivity for food. The code for the food representation is not yet clear. Due to the variability of the food images, their location, image perspective, and most importantly that they recruit areas outside primary visual cortex, it is probable that this is not due to low level features. The lack of consistency between image embedding clusters and voxel activity clusters also suggests that the proposed regions are not due to raw input features.

Attention and/or saliency is not a convincing hypothesis on its own because other categories such as faces also have both attention and saliency, and yet there are shown to recruit a subset of the ventral temporal cortex. These results seem to reliably show there are some regions which are especially activated by images pertaining to food. Future work can use other modalities like MEG to see if there is a modulation of these regions from higher level areas. As they are located in the high level visual system, it is possible at least some of their function is feed-forward or automatic. Further consideration of the effect of individual preferences would help better understand the function of these regions.

Bibliography

[1] Nick Statt. Facebook acquires neural interface startup ctrl-labs for its mind-reading wristband, Sep 2019. URL https://www.theverge.com/2019/9/23/20881032/facebook-ctrl-labs-acquisition-neural-interface-armband-ar-vr-deal. 1

[2] Elon Musk et al. An integrated brain-machine interface platform with thousands of channels. *Journal of medical Internet research*, 21(10):e16194, 2019. 1

[3] Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8(3):293–321, 1992. 1

[4] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982. 1

[5] Aina Puce, Truett Allison, JOHN C Gore, and Gregory McCarthy. Face-sensitive regions in human extrastriate cortex studied by functional mri. *Journal of neurophysiology*, 74(3): 1192–1199, 1995. 1

[6] Nancy Kanwisher. Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, 107(25): 11163–11170, 2010. 1

[7] J Sergent, S Ohta, and B MacDonald. Functional neuroanatomy of face and object processing: A positron emission tomography study. *Brain*, 115:15–36, 1992. 2, 5

[8] N Kanwisher, J McDermott, and M M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci*, 17(11):4302–4311, 1997. 2, 4.1, 5

[9] R Epstein and N Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998. doi: 10.1038/33402. 2, 5

[10] P E Downing, Y Jiang, M Shuman, and N Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001. doi: 10.1126/science.1063414. 2, 5

[11] N Kanwisher. Domain specificity in face perception. *Nat. Neurosci.*, 3(8):759–763, 2000. 2

[12] M J Tarr and I Gauthier. FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nat Neurosci*, 3(8):764–769, 2000. doi: 10.1038/77666.

2

[13] VanderLaan. The first taste is always with the eyes: A meta-analysis on the neural correlates of processing visual food cues. *NeuroImage*, 55(1):296–303, 2011. doi: 10.1016/j.neuroimage.2010.11.055. 2

[14] Ruud van den Bos and Denise de Ridder. Evolved to satisfy our immediate needs: Self-control and the rewarding properties of food. *Appetite*, 47(1):24–29, 2006. 2

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 3.1

[16] Allen. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nat Neurosci*, 25(1):116–126, 2022. doi: 10.1038/s41593-021-00962. 2, 3.1

[17] Emilie L Josephs, Haoyun Zhao, and Talia Konkle. The world within reach: an image database of reach-relevant environments. *Journal of Vision*, 21(7):14–14, 2021. 3.2

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3.4

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3.4

[20] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1993.

[21] Donald Knuth. Knuth: Computers and typesetting. URL `http://www-cs-faculty.stanford.edu/~{}uno/abcde.html`.

[22] Carolijn Ouwehand and Esther K Papies. Eat it or beat it. the differential effects of food temptations on overweight and normal-weight restrained eaters. *Appetite*, 55(1):56–60, 2010.