How the Brain Makes Sense of Natural Scenes

A Major Qualifying Project
submitted to the Faculty of
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
degree of Bachelor of Science
Computer Science and Data Science

By:
Jose Morales, Shivali Mani, Ali McNicholas
Date: April 26th, 2024

Report Submitted to:
Professor Ziming Zhang, Department of Electrical and Computer Engineering

Abstract

Neuroscience and computational modeling have a symbiotic relationship, with discoveries in each field inspiring the other. This paper explores the relationship between visual stimuli from the Natural Scenes Dataset and functional magnetic resonance imaging (fMRI) activity in distinct brain regions of interest (ROI). Valuable information can be extracted from images using components of various pretrained vision models known as feature extractors. This extracted information can be used by neural networks to predict how each ROI will respond to stimuli. Observing the patterns and behaviors, we identified specific regions of interest corresponding to categories identified by a classification model. This study found that YOLOv8n for classification and ResNet50 for feature extraction work best alongside linear regression. We then analyzed the patterns among the categories to identify which classes have similar activations.

Acknowledgment

We would like to express our sincere gratitude to all those who have contributed to the completion of this project. Special thanks to our advisor, Professor Ziming Zhang, for his guidance and support throughout this research project. We are also thankful for the necessary facilities and resources from the Vision, Intelligence, and System Laboratory provided by Professor Ziming Zhang.

Table of Contents

Abstract 1

Acknowledgment 2

Table of Contents 3

Introduction 4

Background 5

- The Occipital Lobe and Vision 5
- Functional Magnetic Resonance Imaging 6
- Natural Scenes Dataset 6
- Regression Models 7
- Neural Networks 8
- Previous Work 10
- Research Objectives 10

Methodologies 12

- Preprocessing 12
- Model Overview 13
- Classification 13
- Grad-CAM 15

Results 16

- Backbone Architectures 16
- Optimizing Regression Models 16
- Model Selection 17
- Comparing Two Subjects: Subject 1 and 5 18
- Comparing Activations 18
- Further Steps 20

Conclusion 21

References 22

Introduction

Understanding how the human brain works is crucial in science and society. Contemporary tools such as neural networks are coming to the forefront when understanding the brain. Technology utilizing functional magnetic resonance imaging (fMRI) examines changes in blood flow to interpret brain activity concerning a specific task, experience, or behavior. This non-invasive technology is widely used in different types of neuroscience research (Whitten, 2012). Findings from these studies advance the understanding of neural responses to various stimuli and tasks. In this paper, we analyze the brain activity from fMRI scans of eight subjects in response to images from the Natural Scenes Dataset (NSD). Using this data, we investigate the correlation between the images and specific fMRI activity while investigating the context of such activations.

Background

The Occipital Lobe and Vision

Vision is the primary sensory input humans use to understand their surroundings, and as such, our brains are coded to treat such stimuli very specifically (Ionta, 2021). Vision starts in the eyes where signals are processed into neural inputs that are sent to the occipital lobe, mainly the visual cortex. Signals are then distributed to sub regions, otherwise known as regions of interest (ROIs). As seen in Figure 1, these regions play different roles in processing information related to places, spatial relationships, motion, and stimulus identification. The stimuli are then sent to the parietal lobe, which allows recognition and adaptation to physical space, and the temporal lobe, which controls memory and assigns the stimuli to the visual stimuli (Debrowski, 2020).

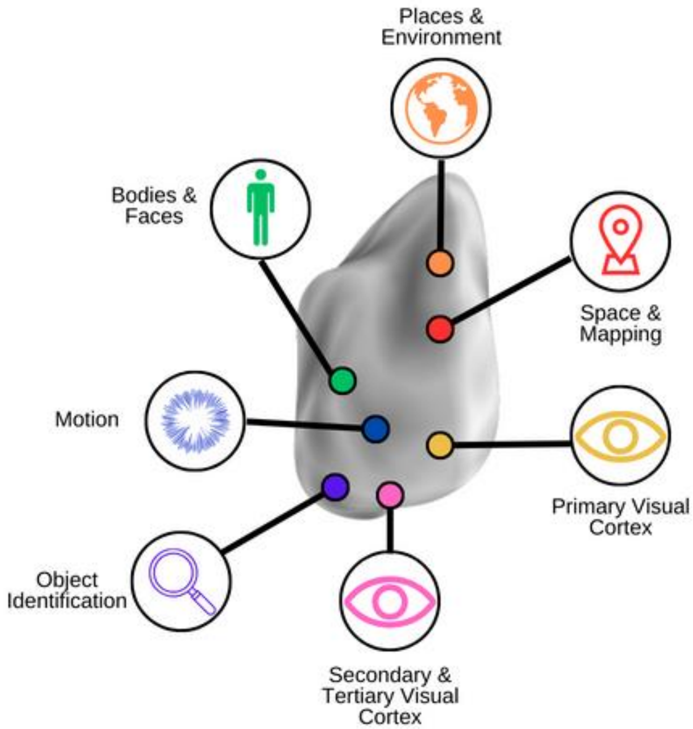

Figure 1: Diagram of how the brain processes images

Functional Magnetic Resonance Imaging

Functional magnetic resonance imaging (fMRI) analyzes changes in blood flow to examine how the brain works regarding a specific task, experience, or behavior (Whitten, 2012). This technology is non-invasive and is widely used in different types of neuroscience research. Findings from these studies aid in the diagnosis, treatment, and prevention of different psychological and neurological disorders. Furthermore, these studies can help us understand different brain activities in response to stimuli.

Natural Scenes Dataset

For this study, we used the Natural Scenes Dataset (NSD), composed of 73,000 images of natural scenes curated from the COCO dataset. The COCO dataset is a benchmark dataset in computer vision known for its diversity and complexity in its images. (Torres, 2024)

Accompanying these images are recorded fMRI activations from eight different subjects. Each subject viewed approximately 10,000 images, 9,000 unique and 1,000 shared among the subjects. The fMRI data was split between the left and right hemispheres of the brain, with a differing number of recorded voxels between the subjects. With the exception of subjects 6 and 8, the left and right hemispheres had 19,004 and 20,544 recorded voxels, respectively. For subjects 6 and 8, due to missing data, the left and right hemispheres had 18,978 and 20,220 recorded voxels, respectively (Allen et al., 2022).

Regression Models

Regression models explore the relationship between independent variables, features, and a dependent variable, or outcome. This study investigates the use of three types of regression models: Linear Regression, Decision Tree Regression, and Multilayer Perceptron (MLP) Regression.

Linear regression predicts a dependent variable using one or more independent variables (Pai, 2021). Multiple linear regression, which we will test in this study, aims to minimize the degree of error between the predictions and the actual values. This involves estimating the best-fitting hyperplane while minimizing the sum of squared differences. A major assumption in this method is that the relationship between the variables is linear.

Decision tree regression utilizes supervised learning to train a model for predicting continuous data (Kotsiantis, 2011). This model is simple and requires little preprocessing of the data. Its primary strategy is to predict the target variable by determining simple rules learned from the input data. The dataset size and complexity are proportional to that of the tree. However, decision trees may not generalize well to new data, which can create biases in certain classes.

A multilayer perceptron is a neural network traditionally used in pattern recognition and interpolation (Choudhury, 2022). When utilized as a regression model, the primary strategy is to minimize the mean squared error of the predicted and ground truth values through an optimization technique such as stochastic gradient descent. The model is trained iteratively, and a regularization term can be added to prevent the model from overfitting to the data. This approach is useful when the data presents a complex relationship that a linear model may not entirely capture.

Neural Networks

Neural Networks are complex deep learning algorithms designed to mimic the decision-making processes of the human brain (Hardesty, 2017). In this study, we will evaluate several networks designed to operate on images: EfficientNet, InceptionNet, ResNet50, ResNet101, VGG19, YOLOv8n, WideResNet50, and WideResNet101. These models will be used as feature extractors, also known as backbones, to extract prominent image data that will be essential in predicting brain activity.

Convolutional Neural Network (CNN) as a primitive neural network utilizes convolutional layers as learnable filters or kernels. This detects patterns where pooling layers will down sample the spatial dimensions to control overfitting and computational complexity. Of which non-linear activations are applied to enable learning complex relationship of the data (IBM, n.d.).

The EfficientNetV2 model is a more recent convolutional network that progressively scales the input image size from a low-resolution image to a higher-resolution image. As the images move to a higher resolution, more data augmentation techniques are applied. This not only decreases the size of the model significantly, but it also performs much quicker than most models while maintaining great classification performance (Tan & Le, 2021).

The InceptionNetV3 model, also known as GoogLeNet, introduces multiple parallel convolutional pathways of different sizes and pooling operations. It can capture extract information from both global and local features, with batch normalization accelerating the training process. It uses global average pooling which facilitates translation-invariant representation and reducing overfitting (Szegedy et al., 2015).

ResNet50 is a neural network that utilizes more than 50 convolutional layers split between weight layers and skip connections. Notably, it uses residual mapping instead of direct learning which allows for easier optimizations compared to other similar types of deep learning algorithms (Tsang, 2019). With experiments on the ImageNet dataset, the study increased accuracy from the increased depth. Furthermore, the analysis showed that these residual functions are closer to the real mappings of the data. ResNet101 is an extension of ResNet50, that includes 101 convolutional layers including the same skip connections (He et al., 2015). WideResNet50 builds upon ResNet50 by increasing the width of the model, the number of neurons in each layer of the network (Zagoruyko & Komodakis, 2016).

The VGG-19 network from the Visual Geometry Group (VGG) uses residual learning with 152 layers to classify different images (Simonyan & Zisserman, 2014). It mainly consists of convolutional layers and max pooling layers stacked on each other. The residual functions present in this model help decrease the degradation problem in many other deep learning algorithms.

YOLO, or You Only Look Once, is a model primarily used to identify objects in a photo or classify the overall image utilizing bounding boxes and class probabilities for multiple objects in a single pass. It is highly efficient and uses low computational resources. Additionally, detection is much simpler than other models, only using a single regression model of image pixels to class probabilities (Redmon et al., 2016). Finally, the YOLOv8n model was trained on the COCO dataset, which proves advantageous for our current study.

Previous Work

Regression models and neural network usage are increasing in neuroscience. Previous work in both fields has been instrumental in this study in predicting fMRI voxels. Haxby et al. (2001) used a nearest-neighbor model with feature extraction to detect patterns in the ventral temporal cortex to images of faces, cats, scissors, and shoes. Norman et al. (2006) also demonstrated that multi-voxel pattern analysis using regression models could detect information in complex patterns. Pereira (2009) also explored fMRI patterns by associating different classes with a specific superclass. Similarly, Han et al. (2015) modeled the memorability of video clips and predicted how memorable they are by utilizing fMRI scans and using audiovisual data. Then, features were derived to calculate the connectivity between brain ROIs from a few brain networks. Each study aimed to find a pattern to answer a question related explicitly to fMRI scans but has yet to attempt to understand how the brain works for a high-performing machine learning prediction. This study focused on advancing the integration of machine learning and neuroscience and aiding the field for further research.

Research Objectives

When tackling this problem, we derived three significant research objectives:

1. Can we accurately predict the fMRI activity associated with each image?

2. Were there distinct activations among different categories of images?
3. How do the backbones and regression models above differ in terms of performance, and which proves most accurate?

Methodologies

Our approach to answering the previously answered research objectives is highlighted in Figure 2. Given the input image, we extract valuable information using one of the backbones mentioned above (Resnet50, VGG19, YOLOv8n) to train a model capable of predicting brain activation in response to the visual stimuli presented in the input image.

Preprocessing

The images in the dataset were resized and preprocessed according to the neural network being used. This often meant that images were resized to 224 x 224 pixels, then converted to a tensor to be compatible with the Pytorch library, and finally normalized across each channel red channel, green channel, and blue channel) with mean 0.485, 0.456, and 0.406, and standard deviation of 0.229, 0.224, and 0.225. These steps were crucial for preparing the data for feature extraction. Additionally, we clipped the fMRI data at the 5th and 95th percentile, extracting any outliers in the dataset. The remaining data was then scaled into values between 0 and 1 using min-max normalization.

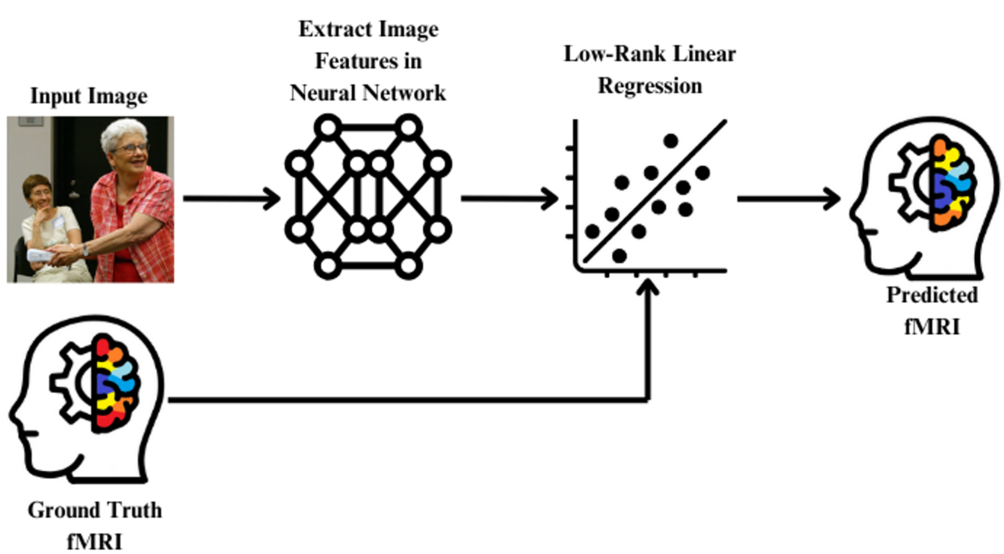

Figure 2: Overview of model.

Model Overview

In determining which configuration resulted in the highest performance, we evaluated different classification models, neural networks, and regression models. First, we tested YOLOv8n and VGG to determine which classifications led to the highest performance given the images available in the dataset. Then, we evaluated the performance of EfficientNet, InceptionNet, ResNet50, ResNet101, VGG19, YOLOv8n, WideResNet50, and WideResNet101 as the backbone, or the images' feature extractor. Once the features have been extracted, we reduced the dimensionality of the resulting data, keeping only the most informative information using Principal Component Analysis (PCA). We experimented with feature dimensions of 8, 16, 32, and 64. For regression model analysis, we evaluated linear regression, decision trees, and multi-layer perception regression using mean squared error (MSE), Pearson correlation, and cosine similarity. We use Pearson correlation as our primary metric in determining the models performance. Finally, we visualized the different fMRI activations to compare the predictions to the actual activity.

Classification

To determine patterns in brain activation between different categories of images, we label each image to allow for comparison using the detected class with the highest confidence score from an object detection model. While we may use the labels provided by the COCO dataset, we found that the COCO dataset's labels were not generalized enough for our goals in this study. The COCO labels reflect smaller detailed items rather than the main object in the photo. We also tested using one of the vision models mentioned above (EfficientNetV2, InceptionNetV3, etc). However, the classes predicted by these models were not ideal for building generalizable labels, with these models predicting over 1000 classes, proving very difficult to map these categories to broader classes. The YOLOv8n worked the best due to the 80 classes it predicts, which can easily be mapped to our list of super classes. Additionally, we found that YOLOv8n detected the most relevant object in a photo rather than a minor detail. Table 1 contains a list of superclasses used, as well as the mapping of each class to its corresponding superclass.

Table 1 Superclass names and their associated class names

Super Class	Class Names
Person	Person
Animal	Bird, Cat, Dog, Horse, Sheep, Cow, Elephant, Bear, Zebra, Giraffe
Toy	Frisbee, Sports Ball, Kite, Teddy Bear
Container	Bottle, Wine Glass, Cup, Bowl
Utensil	Fork, Knife, Spoon
Appliance	TV, Laptop, Mouse, Remote, Keyboard, Cell Phone, Microwave, Oven, Toaster, Sink, Refrigerator, Hair Drier, Toothbrush
Vehicle	Bicycle, Car, Motorcycle, Airplane, Bus, Train, Truck, Boat
Outdoor	Traffic Light, Fire Hydrant, Stop Sign, Parking Meter, Bench
Accessory	Backpack, Umbrella, Handbag, Tie, Suitcase
Food	Banana, Apple, Sandwich, Orange, Broccoli, Carrot, Hot Dog, Pizza, Donut, Cake
Furniture	Chair, Couch, Potted Plant, Bed, Dining Table, Toilet
Indoor	Book, Clock, Vase, Scissors

Grad-CAM

We further investigate the relationship between the visual stimuli in the NSD dataset and the associated fMRI data in the other direction. Specifically, given the predicted fMRI data for a specific region of interest, we aim to visualize the parts of the input image that contribute the most to the prediction. For this, we utilize Grad-CAM, a technique that visualizes the decision-making process for CNN networks, using the gradients computed during the prediction process, particularly those that flow into the last convolutional layer (Selvaraju et al., 2019). The model used in this approach required changes to visualize the regions of the image responsible for predicting fMRI data for a specific ROI. More concretely, using the provided mapping in the NSD dataset, we identified and averaged the recorded voxels within each ROI. We then trained a model using the averaged fMRI data which allowed us to predict the average activation for each ROI instead of each voxel.

Results

Backbone Architectures

Table 2: Neural network results when used as the backbone of image extraction

Backbone	***MSE***	***Corre***
EfficientNetV2	0.49	11.6
InceptionNetV3	0.44	31.7
VGG-19	0.41	37.8
ResNet50	0.41	37.7
Resnet101	0.44	33.8
WideResNet50_2	0.43	35.5
WideResNet101_2	0.42	35.1
Yolov8N	0.44	32.6

The results of our experiments testing various backbones are listed in Table 2. Each of the backbones tested used subjects 1 data and only the results for the left hemisphere are listed here as the values closely match the right hemisphere. VGG-19 and ResNet50 had similar performance with a MSE of 0.41 and correlations of 37.8 and 37.7 respectively.

Optimizing Regression Models

The results of our experiments testing different regression models are listed below in Table 3. We test our results on each subject and report the MSE and Pearson Correlation values. The linear regression model proved to perform the best, both in terms of a lower MSE and a higher Pearson correlation score.

Table 3: MSE and Correlation metrics for Linear Regression (LR), Decision Tree (DT), and Multilayer perceptron (MLP).

Subject	LR MSE	DT MSE	MLP MSE	LR Corre	DT Corre	MLP Corre
1	0.016	0.03	0.019	40.4	13.7	29.7
2	0.015	0.03	0.018	42	15	33
3	0.014	0.031	0.018	35	11	24.5
4	0.014	0.03	0.017	34	11.5	23.5
5	0.015	0.03	0.017	46	20	38
6	0.014	0.03	0.017	34	10	23.5
7	0.015	0.03	0.018	33	11	24.7
8	0.013	0.028	0.016	26	8	15

Model Selection

We chose to use ResNet50 with linear regression as our final model. Though both VGG-19 and ResNet50 have the same MSE and a difference of 0.1 in correlation, the choice to use Resnet50 was due to the difference in model size, with ResNet50 containing 25.6 million parameters (He et al., 2015) and VGG-19 143.7 million parameters (Simonyan & Zisserman, 2015).

Comparing Two Subjects: Subject 1 and 5

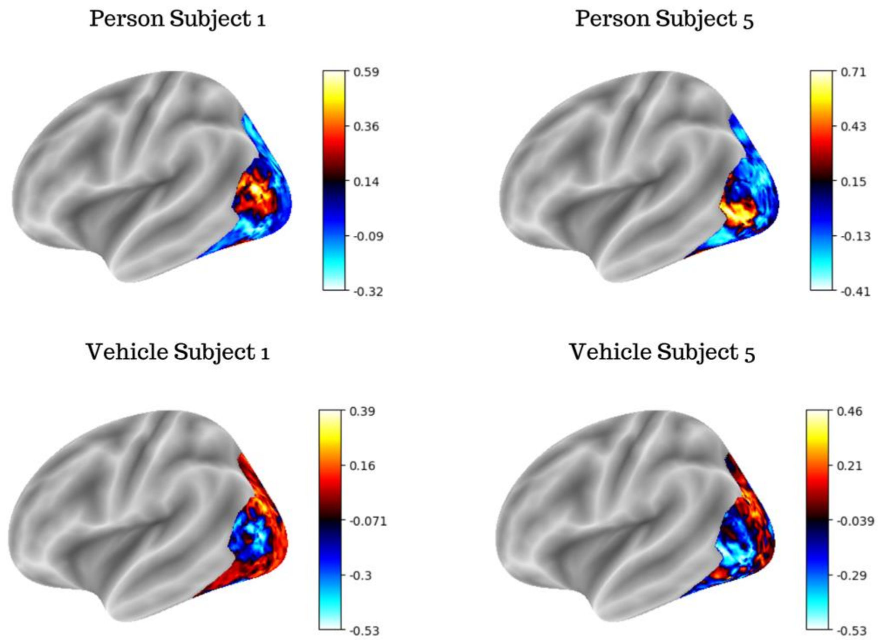

Figure 3: Left hemisphere visualizations on the food and person superclass.

To address our question if the fMRI results would be similar among patients, Figure 3 displays two super classes from different patients. Subjects 1 and 5 have similar activation patterns when they observe people or vehicles but are still unique. These results support the previous claims that different subjects could have different fMRI activations (Bassett et al., 2008). Because of this finding, we decided to train our model individually on the subjects.

Comparing Activations

When observing activations, we found patterns among the different visualizations in the super classes. In Figure 4, we discovered similarities among living creatures, humans and animals, food-related items, food and utensils, and inanimate objects. When looking further into the activation, we noticed that the living creatures and non-food-related inanimate objects were in contrast, specifically in the extrastriate body area (EBA) region. When inanimate objects are viewed, regions such as the occipital place area (OPA), parahippocampal place area (PPA), and retrosplenial cortex (RSC) are activated due to their nature in detecting scenes or objects. Additionally, there is a distinct activation in the lower lateral and the upper ventral area in the left hemisphere when food items are viewed. These activations confirm prior research indicating that different brain regions respond to different stimuli. Furthermore, this type of modeling allows researchers to understand new associations the brain makes to classify different visual stimuli.

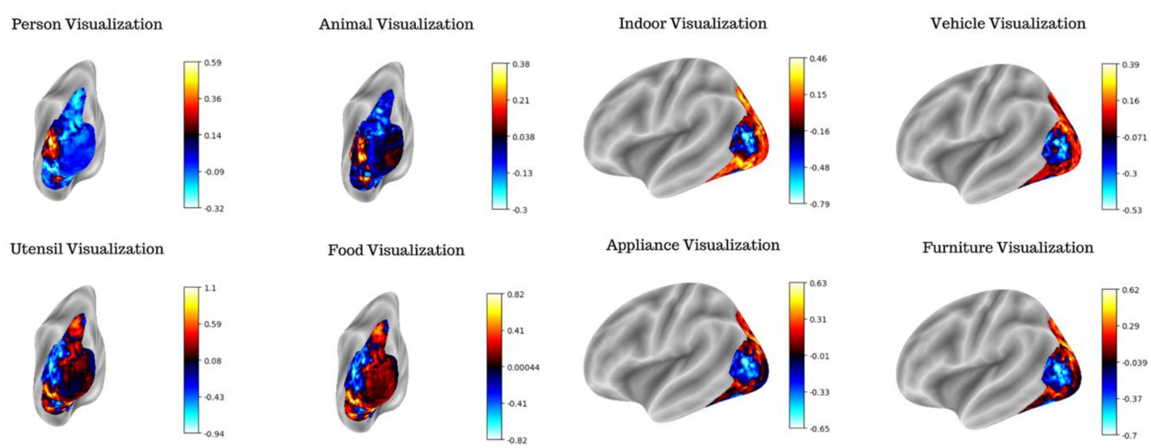

Figure 4: Brain activations of the person, animal, food, utensil, indoor, vehicle, appliance, and furniture categories

Additionally, we calculated the cosine similarity among the super classes. As seen in Table 5, the cosine similarity of the predicted values and real values are high, which shows that they are like the real results. We only included values with a count greater than 10 to ensure we had enough samples to summarize the predictions accurately.

Table 5: Cosine Similarities for Different Categories for different subjects.

Class	**Subject 1**	**Subject 2**	**Subject 3**	**Subject 4**	**Subject 5**	**Subject 6**	**Subject 7**	**Subject 8**
Person	0.98	0.98	0.98	0.97	0.99	0.98	0.98	0.96
Animal	0.95	0.96	0.94	0.86	0.97	0.95	0.93	0.94
Vehicle	0.96	0.97	0.94	0.96	0.98	0.96	0.97	0.92
Outdoor	0.78	0.8	0.63	0.63	0.88	0.77	0.75	0.68
Accessory	0.57	0.5	0.52	0.49	0.28	0.62	0.43	0.15
Toy	0.92	0.9	0.7	0.81	0.87	0.83	0.88	0.73
Container	0.93	0.93	0.91	0.88	0.97	0.92	0.93	0.92
Utensil	0.87	0.88	0.93	0.86	0.92	0.76	0.78	0.81
Food	0.98	0.98	0.98	0.97	0.98	0.95	0.97	0.96
Furniture	0.97	0.97	0.96	0.97	0.98	0.94	0.95	0.96
Appliance	0.95	0.98	0.95	0.95	0.97	0.95	0.94	0.92
Indoor	0.93	0.93	0.9	0.89	0.95	0.87	0.9	0.88

Further Steps

While this study delved into various neural networks and regression models, further testing could enhance results. Moreover, employing a more diverse classification model could better identify images and categorize them into different super classes. Furthermore, conducting a follow-up study using a custom model trained specifically on the data, rather than a pretrained one, may yield different outcomes.

Conclusion

In this study, we investigate how different natural scenes activate distinct ROIs in the brain with fMRI data. Utilizing deep learning models, such as YOLOv8n and VGG network, we used linear regression to predict fMRI voxels and identify different brain activity patterns. From this, we found distinct activations from the different categories.

We used MSE, R-Squared, and cosine similarity to evaluate the performance. While the MSE scores were ideal, the R-squared scores could be improved. The cosine similarity allowed us to identify the high similarities in the categories with the fMRI data.

References

1. Allen, E.J., St-Yves, G., Wu, Y. *et al.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat Neurosci* **25**, 116–126 (2022). https://doi.org/10.1038/s41593-021-00962-x

2. Bassett, S. S., Scott, D. J., Honeycutt, N. A., Verduzco, G., Yassa, M. A., & Cuzzocreo, J. L. (2008, June 20). *Effect of handedness on fMRI activation in the medial temporal lobe during an auditory verbal memory task*. Wiley Online Library. https://onlinelibrary.wiley.com/

3. Choudhury, K. (2022, August 4). *Deep Neural Multilayer Perceptron (MLP) with Scikit-Learn*. Medium. https://towardsdatascience.com/deep-neural-multilayer-perceptron-mlp-with-scikit-learn-2698e77155e

4. Debrowski, A. (2020, March 8). *How does the brain control eyesight?*. All About Vision. https://www.allaboutvision.com/resources/part-of-the-brain-controls-vision/A

5. Hardesty, L. (2017, April 14). *Explained: Neural networks*. MIT News | Massachusetts Institute of Technology. https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

6. Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, 293(5539), 2425-2430.

7. He, K., Zhang, X., Ren, S., & Sun, J. (2015, December 10). Deep residual learning for image recognition. arXiv.org. https://arxiv.org/abs/1512.03385

8. Ionta, S. (2021, May 6). *Visual neuropsychology in development: Anatomo-functional brain mechanisms of action/perception binding in health and disease*. Frontiers. https://doi.org/10.3389/fnhum.2021.689912

9. Han J, Chen C, Shao L, Hu X, Han J, Liu T. Learning Computational Models of Video Memorability from fMRI Brain Imaging. IEEE Trans Cybern. 2015 Aug;45(8):1692-703. doi: 10.1109/TCYB.2014.2358647

10. Kotsiantis, S. B. (2011, June 29). *Decision trees: A recent Overview - Artificial Intelligence Review*. SpringerLink. https://link.springer.com/article/10.1007/s10462-011-9272-4

11. Norman, K., Polyn, S. M., Detre, G., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences, 10(9), 424-430.

12. Pai, A. (2021, October 6). *Multiple linear regression*. Machine Learning Works. https://www.machinelearningworks.com/tutorials/multiple-linear-regression

13. Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. NeuroImage, 45(1 Suppl), S199-S209. https://doi.org/10.1016/j.neuroimage.2008.11.007

14. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016, May 9). *You only look once: Unified, real-time object detection*. arXiv.org. https://arxiv.org/abs/1506.02640

15. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019, December 3). Grad-cam: Visual explanations from deep networks via gradient-based localization. arXiv.org. https://arxiv.org/abs/1610.02391

16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. *CoRR*, *abs/1512.00567*. Retrieved from http://arxiv.org/abs/1512.00567

17. Simonyan, K., & Zisserman, A. (2015, April 10). *Very deep convolutional networks for large-scale image recognition*. arXiv.org. https://arxiv.org/abs/1409.1556

18. Tan, M., & Le, Q. V. (2021, June 23). EFFICIENTNETV2: Smaller models and faster training. arXiv.org. https://doi.org/10.48550/arXiv.2104.00298

19. Torres, J., & Austen, J. T. J. (2024, March 8). *Yolov8 coco dataset: Powerful combination for object detection*. YOLOv8. https://yolov8.org/yolov8-coco-dataset/

20. Tsang, S.-H. (2019, March 20). *Review: Resnet - winner of ILSVRC 2015 (Image Classification, localization, detection)*. Medium. https://towardsdatascience.com/review-resnet-winner-of-ilsvrc-2015-image-classification-localization-detection-e39402bfa5d8

21. *What are convolutional neural networks?*. IBM. (n.d.). https://www.ibm.com/topics/convolutional-neural-networks

22. Whitten, L. A. (2012). Functional magnetic resonance imaging (fMRI): An invaluable tool in translational neuroscience. RTI Press publication No. OP-0010-1212. Research Triangle Park, NC: RTI Press. Retrieved from http://www.rti.org/rtipress.

23. Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. ArXiv.org. https://doi.org/10.48550/arXiv.1605.07146