

IMG2BRAIN: PREDICTING THE NEURAL RESPONSES TO VISUAL STIMULI OF NATURALISTIC SCENES USING MACHINE LEARNING

CAPSTONE PROJECT REPORT
MACHINE LEARNING AND MULTIVARIATE STATISTICS - MSB1011
ACADEMIC YEAR: 2022-2023

 **Sebastian Ayala Ruano**
Faculty of Science and Engineering
Maastricht University
Student ID: i6314501
`s.ayalaruano@student.maastrichtuniversity.nl`

June 9, 2023

ABSTRACT

Computational models are essential tools for interpreting the large datasets produced by fMRI experiments. Visual encoding models based on fMRI data employ algorithms that transform image pixels into model features and map these features to brain activity. For this project, visual encoding models were created to predict the neural responses to visual stimuli of naturalistic images. Four pre-trained CNNs (AlexNet, VGG16, ResNet50, and InceptionV3) were tested to extract the features of the images and select the best feature representation to build the encoding models. Then, six machine learning algorithms (linear regression - base model, ridge regression, lasso regression, elasticnet regression, k-nearest neighbors regressor, and decision tree regressor) were trained to predict the brain activity of the voxels from the feature representation of images. The best model was the lasso regression with an encoding accuracy of 0.2417 on the validation set. Although the performance output of the best model was low, it is a starting point to build upon. All underlying data and code are accessible through GitHub (github.com/sayalaruano/img2brain) under the MIT and CC0 licenses and archived on Zenodo ([10.5281/zenodo.7979730](https://doi.org/10.5281/zenodo.7979730)).

Keywords Visual encoding models · Machine learning · Deep learning

1 Introduction

Vision is one of the main sensory pathways that enable living organisms to perceive external stimuli and interpret the world. The human visual system (HVS) involves a complex interplay between the eyes, the brain, and multiple neural pathways [1]. When we observe something, our eyes detect patterns of light, which are then processed and transmitted to the brain for interpretation. Understanding the perceptual, cognitive, and neural processes involved in visual perception is a thriving field of current research. To study the HVS, investigators employ various experimental methods such as psychophysics, eye tracking, and neuroimaging techniques like electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) [2].

fMRI has been extensively utilized due to its non-invasive nature, high spatial and temporal resolution capabilities, and the ability to study brain activity and its relationship with sensory perception. This technique uses magnetic resonance imaging to measure the blood-oxygen-level-dependent (BOLD) variation induced by neuronal activity. By detecting these changes, fMRI maps the functional activation of specific brain regions to specific behaviors, providing valuable insights into the underlying mechanisms of the HVS [1].

The data acquired through fMRI experiments are stored in the form of voxels, which are three-dimensional units representing tiny volume elements. Each voxel encompasses millions of brain cells, collectively forming areas with

different functional properties, known as regions of interest (ROIs). These voxels capture neural activity and translate it into continuous values, reflecting the level of activation within specific ROIs associated with different stimuli [3].

Computational models are essential tools for interpreting the large datasets produced by fMRI experiments. From a computational standpoint, the HVS can be conceptualized as a sequence of encoding and decoding processes. The encoding process aims to predict brain activities triggered by visual stimuli from the external environment. In contrast, the decoding process analyzes brain activities to retrieve the associated visual stimuli. It is worth noting that after constructing an encoding model, the derivation of the decoding model can be achieved through Bayesian inference [4]. Figure 1 shows a representation of the encoding and decoding processes for the HVS.

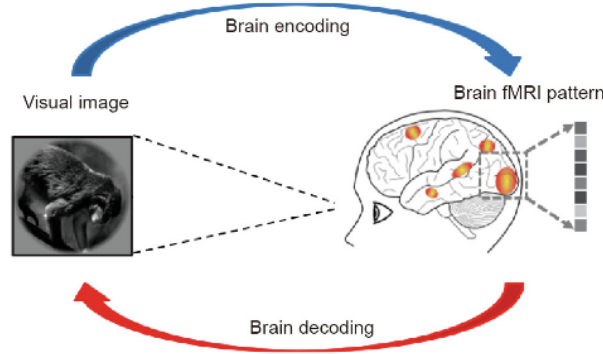


Figure 1: Brain encoding and decoding in fMRI. Obtained from [5]

Visual encoding models based on fMRI data employ algorithms that transform image pixels into model features and map these features to brain activity. Therefore, these models consist of three spaces (the input stimulus space, the feature space, and the brain activity space) and two in-between mappings. The initial mapping between the stimulus and the feature spaces (feature extraction) is commonly nonlinear. On the other hand, the mapping between the feature and the brain activity spaces tends to be linear, as it simplifies the biological interpretation of the results [6]. Figure 2 depicts a scheme of the visual encoding models.

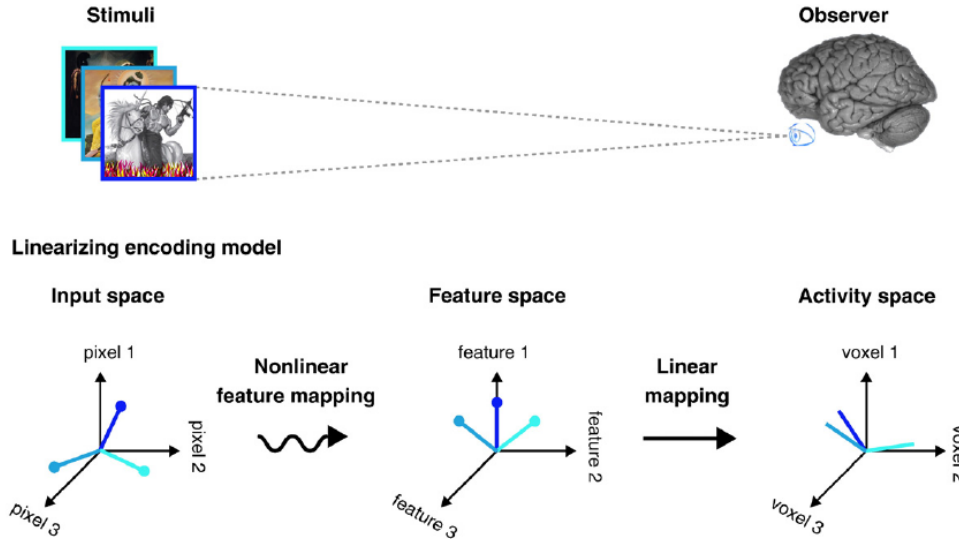


Figure 2: Linearizing visual encoding models. These models are composed of three spaces (the input stimulus space, the feature space, and the brain activity space) and two in-between mappings. Obtained from [4]

In this project, visual encoding models were developed to predict the neural responses to visual stimuli of naturalistic images. The details about the dataset are explained in section 2. First, I tested 4 pre-trained CNNs (AlexNet, VGG16, ResNet50, and InceptionV3) to extract the features of the images and selected the best feature representation to build the encoding models. Then, I trained 6 different machine learning (ML) algorithms (linear regression - base model, ridge regression, lasso regression, elasticnet regression, k-nearest neighbors regressor, and decision tree regressor) to predict

the brain activity of the voxels from the feature representation of the images on the training partition. The models were evaluated with the correlation between the predicted and actual brain activity of the voxels on the validation set. Finally, I selected the best model and predicted the brain activity of the voxels from the images on the test partition. Figure 3 summarizes all the steps for the visual encoding models developed in this project.

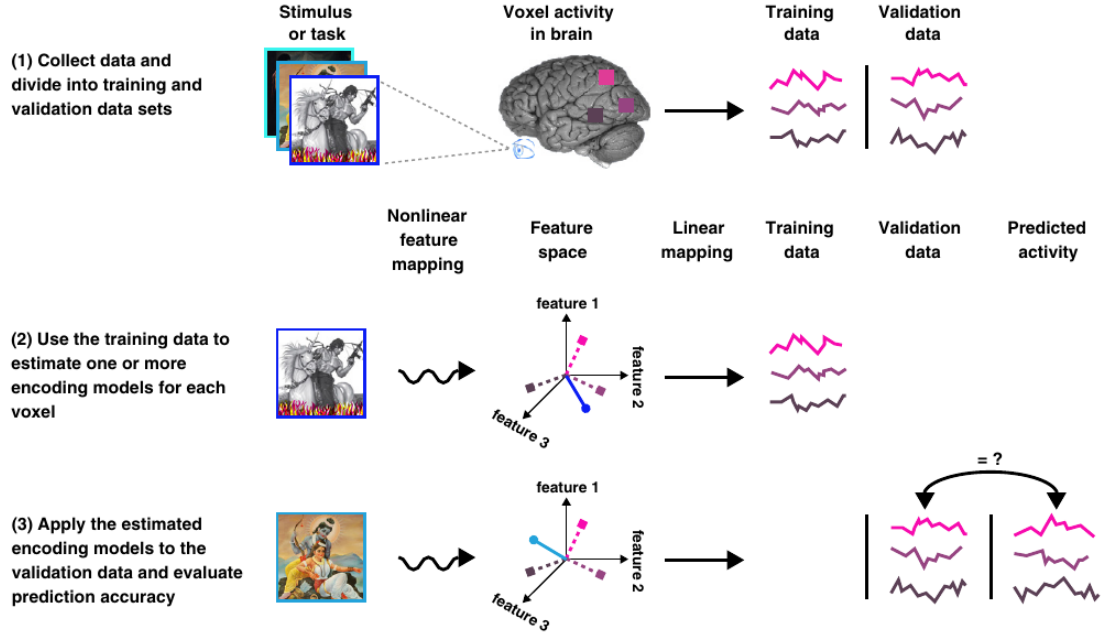


Figure 3: Summary of the main steps for the creation of the visual encoding models. Obtained from [4]

2 Methods

2.1 Dataset

2.1.1 Description

The data for this project is part of the Natural Scenes Dataset (NSD), a massive dataset of 7T fMRI responses to images of natural scenes coming from the COCO dataset [7]. The training dataset consists of brain responses measured at 10,000 brain locations (voxels) to 8857 images (in jpg format) for one subject. The 10,000 voxels are distributed around the visual pathway and may encode perceptual and semantic features in different proportions. The test dataset comprises 984 images (in jpg format), and the goal is to predict the brain responses to these images. The dataset is freely accessible through Zenodo with the following DOI: 10.5281/zenodo.7979730.

2.1.2 Training-validation-test split

The training dataset was split into training and validation partitions with an 80/20 ratio. The training partition was used to train the models, and the validation partition was used to evaluate the models. The test dataset was used to make predictions with the best model on unseen data.

2.2 Feature engineering

Owing to the elevated dimensionality of image feature representation when employing raw pixel values (i.e., original images with a size of 425x425 and 3 channels (RGB), resulting in a feature representation of $425 \times 425 \times 3 = 541,875$ features), an alternative strategy was performed. This involved leveraging representations obtained from pre-trained Convolutional Neural Networks (CNNs) to achieve a lower-dimensional representation of the images. In this study, different layers of four pre-trained CNN models, namely AlexNet, VGG16, ResNet50, and InceptionV3, were evaluated. These models can be accessed through the torchvision Python package [8].

The reduced feature representations of the images were acquired by forwarding the images through the pre-trained CNNs and extracting the output from the designated layer. Despite this, the resulting feature vectors retained a considerable

dimensionality, prompting the utilization of Principal Component Analysis (PCA) to derive a set of 30 features per image. To accomplish this, PCA was fitted on the feature vectors of the training images and subsequently employed to downsample the feature vectors of the training, validation, and test partitions. Due to limitations in available memory (RAM), the PCA computation was executed in batches using the IncrementalPCA class from the scikit-learn Python library [9].

The evaluation of the optimal feature representation was conducted by training a basic linear regression model to predict the brain activity of the voxels using the feature representation of the images. The training set was employed to train the linear regression model, which was subsequently evaluated using the validation set. The best feature representation was determined based on the highest encoding accuracy observed on the validation set. Figure 4 shows a diagram of the feature engineering stage.

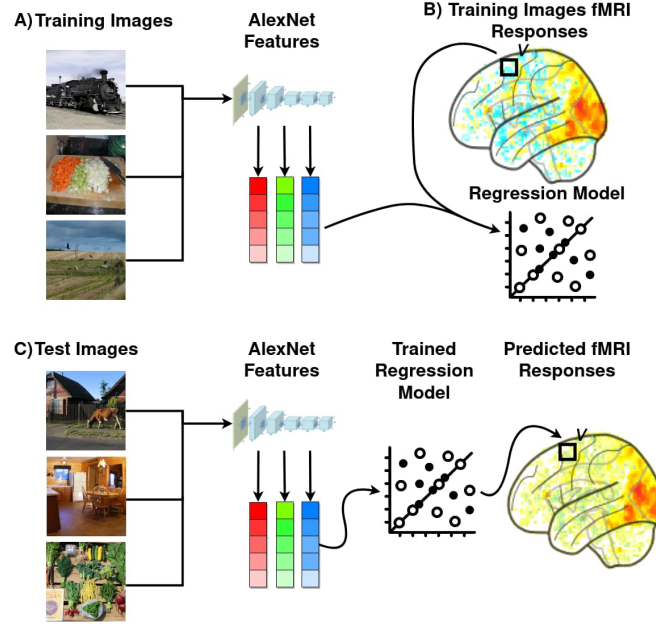


Figure 4: Diagram of the feature engineering stage. Obtained from [10]

Given the GPU RAM requirements of the pre-trained models (at least 10 GB), this analysis was conducted on a GPU-powered Colab platform.

2.3 Machine learning models

Initially, a baseline linear regression model was constructed. Subsequently, three regularized linear regression models, namely ridge, lasso, and elastic net, were developed using a 3-fold cross-validation to identify the optimal hyperparameters. In addition, two non-linear regression models, specifically the k-nearest neighbors regressor and decision tree regressor, were employed.

2.3.1 Training

The learning task for the models was a multi-output regression, where the feature representation of the images served as the input, and the brain activity of all voxels was the output. Each regressor independently mapped the feature space extracted from the CNNs to each voxel, resulting in separate visual encoding models for each voxel. Consequently, every model trained with this dataset encompassed 10,000 independent regression models, each characterized by n coefficients (representing the number of features). All the models and calculations for this section were done with the scikit-learn Python library [9].

2.3.2 Model evaluation

The models were evaluated based on their performance using the validation set. The performance metric to assess the models was the median pearson correlation coefficient between the predicted and actual brain activity across all the

voxels, which will be called the encoding accuracy in the following parts of this report. The top-performing model was the one with the highest encoding accuracy.

3 Results

3.1 Feature engineering

Table 1 shows the results for the encoding accuracy of voxel-wise linear regression models trained with feature representations obtained from different layers of the AlexNet, VGG16, ResNet50, and InceptionV3 CNNs. Intermediate and final layers were prioritized because previous studies reported them as being the most successful for feature representation tasks. The models with the highest encoding accuracy were obtained with layers from AlexNet, and the best feature representation was the intermediate convolutional layer *features.12*.

Table 1: Encoding accuracy of voxel-wise linear regression models trained with feature representations obtained from different layers of various CNNs

Pre-trained CNN ^a layer	Encoding accuracy
AlexNet-features.12	0.205
AlexNet-classifier.6	0.204
AlexNet-classifier.2	0.203
AlexNet-classifier.5	0.203
VGG16-classifier.5	0.198
AlexNet-features.9	0.196
VGG16-features.30	0.195
AlexNet-features.5	0.195
AlexNet-features.7	0.194
VGG16-classifier.6	0.192
AlexNet-features.2	0.2
VGG16-features.20	0.171
ResNet50-layer1.1.conv1	0.171
ResNet50-avgpool	0.158
InceptionV3-Mixed-6c.branch7x7-2.conv	0.158
InceptionV3-Mixed-5c.branch1x1.conv	0.150
InceptionV3-Mixed-6b.branch1x1.conv	0.149
InceptionV3-Conv2d-4a-3x3.conv	0.142
InceptionV3-Mixed-7c.branch-pool.conv	0.140
ResNet50-layer2.3.conv2	0.135
ResNet50-layer3.5.conv3	0.127
ResNet50-layer4.2.conv3	0.121
InceptionV3-avgpool	0.119

^a Convolutional neural network

3.2 Machine learning models

Table 2 presents the encoding accuracy of different voxel-wise machine learning models trained with the best feature representation. These values were obtained by evaluating the models on the validation partition of the dataset. According to these results, the best model was the Lasso regression with an alpha of 0.01 and the default maximum number of iterations set to 1000.

The encoding accuracy values of Table 2 summarize the performance of the models across all voxels, but a general outlook of the distribution from these values provides a more detailed interpretation of the results. Figure 5 shows that the distribution of encoding accuracy values across all voxels for the baseline and regularized linear models are right-skewed with a heavy tail around 0-4-0.7. This observation indicates that the models made accurate predictions on a subset of voxels. Furthermore, as expected from the results of Table 2, the models do not accurately predict the neural responses for a significant portion of voxels. Because of the lack of ROIs information for the brain activity data, it was not possible to identify the visual areas associated with high or low predictions.

Table 2: Encoding accuracy of voxel-wise machine learning models trained with the best feature representation (features.12 from AlexNet). The evaluation was done on the validation partition.

Machine learning model	Encoding accuracy
Lasso-alpha0.01	0.2417
ElasticNet-alpha0.001	0.2415
Ridge-alpha1.0	0.2412
Linear-reg-base-model	0.2402
KNeighborsRegressor	0.1021
DecisionTreeRegressor	0.0382

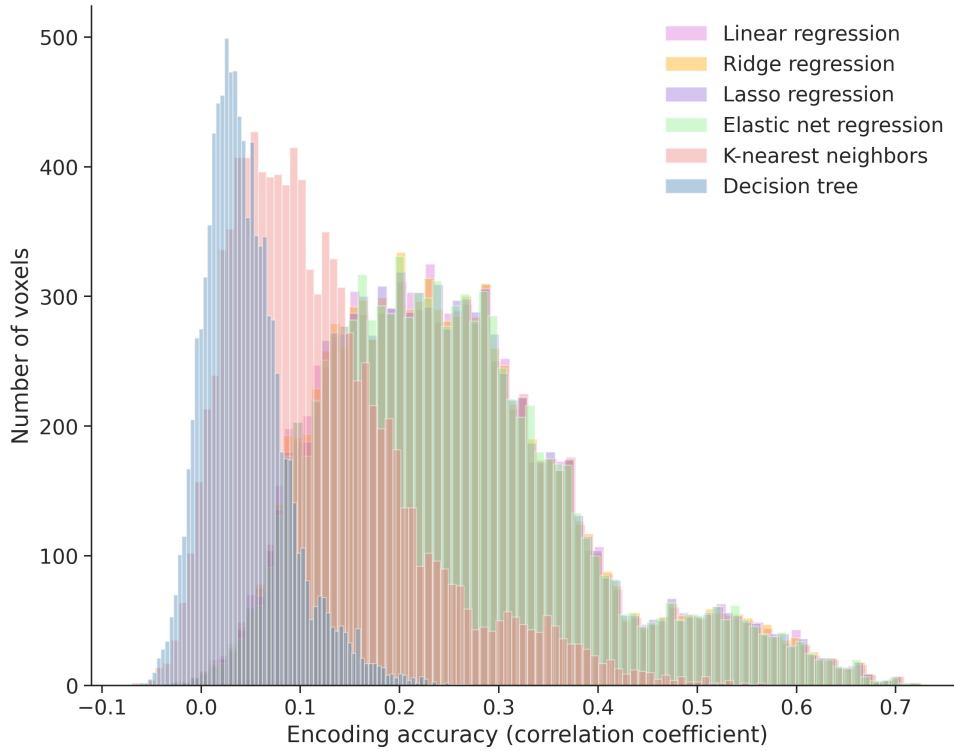


Figure 5: Histograms of the encoding accuracy for the machine learning models across all the voxels. These models were trained to predict the neural responses to visual stimuli of naturalistic images.

As a proof of concept to verify highly and lowly accurate estimations of the best visual encoding model, the predicted and actual BOLD variation induced by neuronal activity for the top-one voxel and the voxel with the lowest encoding accuracy were plotted (Figure 6). The BOLD predictions for the top-one voxel presented a high superposition with the actual BOLD values (Figure 6A), and they were positively correlated (Figure 6B). Conversely, the prediction and actual BOLD values for the voxel with the lowest encoding accuracy did not have an overlap nor a correlation pattern (Figure 6C and D).

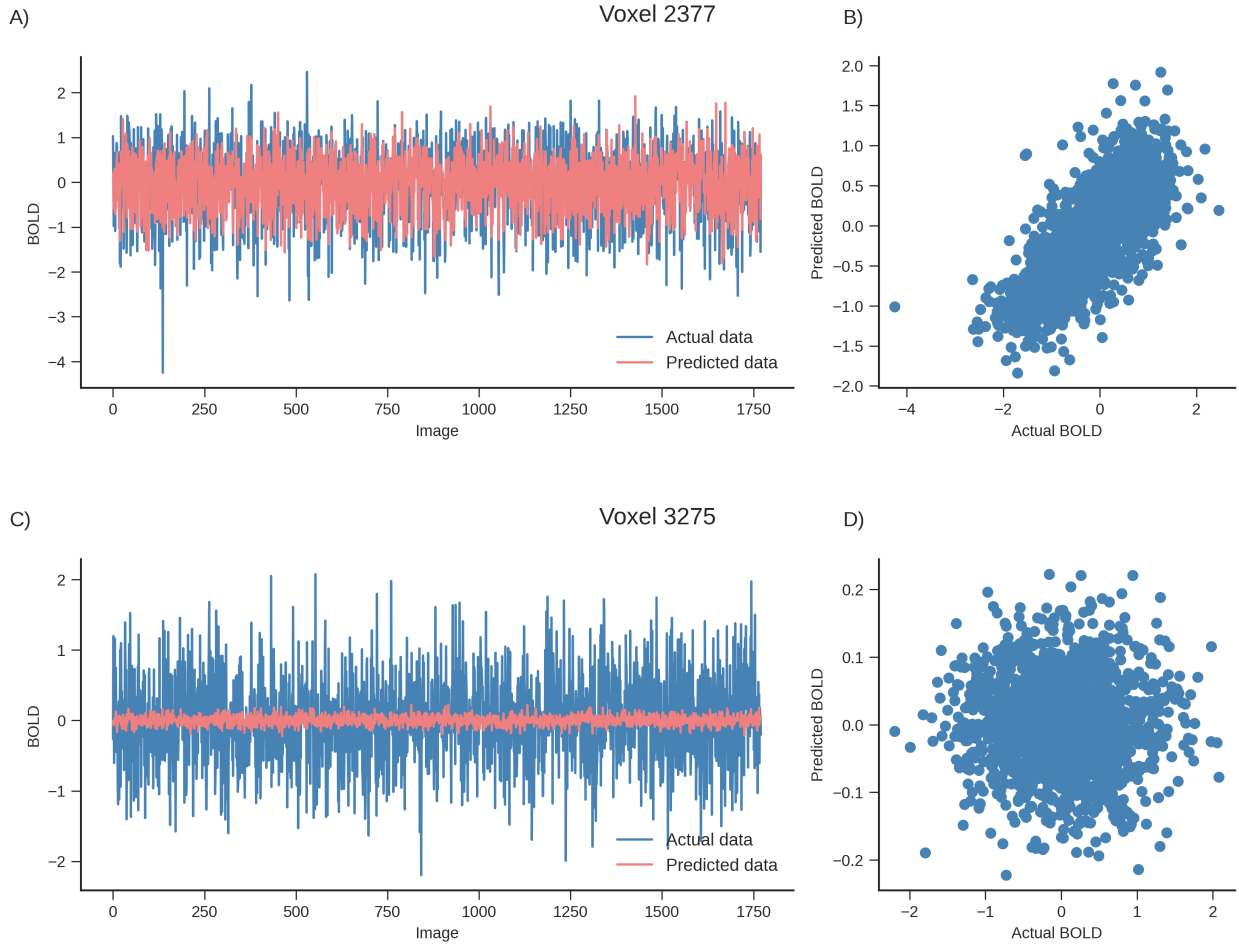


Figure 6: Predicted and actual blood-oxygen-level-dependent (BOLD) variation induced by neuronal activity for the top one voxel (A-B) and the voxel with the lowest correlation coefficient (C-D) from the best visual encoding model (Lasso regression).

4 Discussion

One of the main challenges for this project was the high dimensionality of the dataset in terms of the number of features available for each training instance. Employing all pixels from the images was infeasible because of the computational limitations to handle so many parameters. Therefore, the feature representation from the images was reduced using pre-trained CNNs and PCA. The intermediate convolutional layer *features.12* of Alexnet was the best feature extractor. The explanation for this output can be attributed to the fact that intermediate layers of CNNs have a balance for capturing low-level details (as early layers) and high-level (as end layers) patterns from the images, allowing them to extract meaningful feature representations [11].

After obtaining the feature representation of the images, it was necessary to perform PCA to select the principal components that captured most of the variance of the feature space. The feature engineering steps to reduce the dimensionality of the feature space from the images could be one of the reasons for the low encoding accuracy of the models. By applying these processes, a large portion of the information from the images is lost, which can downgrade

the prediction capacity of the models. Therefore, increasing the number of features in the representation of images could be one potential improvement for further work [11]. Moreover, different dimensionality reduction techniques (e.g., t-SNE, UMAP, autoencoders, etc) could be tested to look for better results. Another area for further research could be the exploration of additional feature representations of the images by using distinct layers from the CNNs or new pre-trained CNNs.

The next part for potential improvement could be the modeling section. The best model was the Lasso regression with an encoding accuracy of 0.2417 on the validation set. Some potential tasks for further development could be a more exhaustive hyperparameter exploration of the ML algorithms, and testing more complex voxelwise encoding models such as random forest or neural networks (the available computational resources were not enough to train these algorithms). One promising approach is applying transfer learning to train a voxelwise encoding CNN. In this project, pre-trained CNNs were used to extract the features of the images, but it would be possible to train a CNN to predict the brain activity of the voxels from the images.

Furthermore, the learning task is something that can be changed. Instead of creating voxel-wise encoding models to predict the responses of each voxel, it is possible to use ROI-wise encoding models to predict the responses of all voxels for each ROI. A previous study created ROI-wise encoding models by extracting the features from a pre-trained CNN and applied transfer learning to train a CNN that predicted the responses of all voxels in each ROI [6]. Finally, another way to improve the models could be to augment the data [3]. The dataset of this project was only part of the Natural Scenes Dataset [7], which contains data for more patients and images.

Due to the lack of information about ROIs associated with the voxels, it was not possible to explore the biological implications of the predictions. Nonetheless, it was shown that the model had high encoding accuracy values for some voxels (Figures 5 and 6). Previous research that created voxel-wise encoding models to understand how visual areas decode mental images of remembered scenes found a similar distribution for encoding accuracy. The authors of this article investigated the predictions for specific ROIs with high encoding accuracy [12]. In consequence, the model developed in this project could deliver accurate results to predict the brain activity of certain ROIs, so further research should focus on this aspect.

Another field for future development could be the use of the encoding models created in this project to derive decoding models and apply them to decode features. These models could link brain activity data to visual stimuli. Decoding models could be derived using Bayes' theorem to invert the linear mapping, and they may be applied to verify the results of the encoding algorithms [5].

As a closing thought, the application of machine learning techniques into neuroscience is a promising avenue to investigate the brain in a data-driven manner. In addition, artificial intelligence may benefit from this synergy to develop computational systems inspired by neurological knowledge [10].

5 Conclusions

For this project, visual encoding models were created to predict the neural responses to visual stimuli of naturalistic images. I tried four pre-trained CNNs (AlexNet, VGG16, ResNet50, and InceptionV3) to extract the features of the images and selected the best feature representation to build the encoding models. Then, I trained six different ML algorithms (linear regression - base model, ridge regression, lasso regression, elasticnet regression, k-nearest neighbors regressor, and decision tree regressor) to predict the brain activity of the voxels from the feature representation of the images on the training partition. The best model was the lasso regression with an encoding accuracy of 0.2417 on the validation set. The best hyperparameters of the model were $\alpha=0.01$ and the default $\text{max-iter}=1000$. This model was trained with the feature representation of the images obtained from the layer features.12 of the AlexNet CNN, which were reduced to 100 features using PCA.

Although the performance output of the best model was low, it is a starting point to build upon. There were several limitations in this project that could be improved. The main limitation was the lack of computational resources to test more pre-trained CNNs to extract the features of the images, increase the number of features in the representation of the images, train more complex machine learning models, and explore more hyperparameters of these algorithms. Also, the learning task was challenging because of the high dimensionality of the feature representation of the images and the multi-output regression task.

References

- [1] Kalanit Grill-Spector and Rafael Malach. The Human Visual Cortex. *Annual Review of Neuroscience*, 27(1):649–677, 2004. [_eprint: https://doi.org/10.1146/annurev.neuro.27.070203.144220](https://doi.org/10.1146/annurev.neuro.27.070203.144220).
- [2] Mark W. Greenlee and Peter U. Tse. Functional Neuroanatomy of the Human Visual System: A Review of Functional MRI Studies. In Birgit Lorenz and Francois-Xavier Borruat, editors, *Pediatric Ophthalmology, Neuro-Ophthalmology, Genetics*, Essentials in Ophthalmology, pages 119–138. Springer, Berlin, Heidelberg, 2008.
- [3] Jonathan D. Cohen, Nathaniel Daw, Barbara Engelhardt, Uri Hasson, Kai Li, Yael Niv, Kenneth A. Norman, Jonathan Pillow, Peter J. Ramadge, Nicholas B. Turk-Browne, and Theodore L. Willke. Computational approaches to fMRI analysis. *Nature Neuroscience*, 20(3):304–313, March 2017. Number: 3 Publisher: Nature Publishing Group.
- [4] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, May 2011.
- [5] Changde Du, Jinpeng Li, Lijie Huang, and Huiguang He. Brain Encoding and Decoding in fMRI with Bidirectional Deep Generative Models. *Engineering*, 5(5):948–953, October 2019.
- [6] Chi Zhang, Kai Qiao, Linyuan Wang, Li Tong, Guoen Hu, Ru-Yuan Zhang, and Bin Yan. A visual encoding model based on deep neural networks and transfer learning for brain activity measured by functional magnetic resonance imaging. *Journal of Neuroscience Methods*, 325:108318, September 2019.
- [7] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. Number: 1 Publisher: Nature Publishing Group.
- [8] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, MM '10, pages 1485–1488, New York, NY, USA, October 2010. Association for Computing Machinery.
- [9] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.
- [10] A. T. Gifford, B. Lahner, S. Saba-Sadiya, M. G. Vilas, A. Lascelles, A. Oliva, K. Kay, G. Roig, and R. M. Cichy. The Algonauts Project 2023 Challenge: How the Human Brain Makes Sense of Natural Scenes, January 2023. [arXiv:2301.03198 \[cs, q-bio\]](https://arxiv.org/abs/2301.03198).
- [11] Manjunath Jogin, Mohana, M S Madhulika, G D Divya, R K Meghana, and S Apoorva. Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning. In *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 2319–2323, May 2018.
- [12] Thomas Naselaris, Cheryl A. Olfman, Dustin E. Stansbury, Kamil Ugurbil, and Jack L. Gallant. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage*, 105:215–228, January 2015.