Life Sciences: Predicting Image Categories using Brain Decoding

Charles Akin-David (aakindav@stanford.edu)
Aarush Selvan (aselvan@stanford.edu)
Minymoh Anelone (manelone@stanford.edu)

Final Report

Introduction

The brain is the final frontier in the study of the human anatomy and can be researched via many different disciplines. One approach - brain decoding - involves the application of brain imaging techniques along with computational/statistical analysis in order to read information from the human brain and analyze it. This in turn can lead to new insights into how the brain works or help develop new technologies. For this project, we worked under the guidance of Professor Justin Gardner in the Psychology Department, to better understand how we can decode human visual object recognition directly from the brain. The Gardner lab at Stanford seeks to understand how neural activity in the human cortex creates our sense of visual perception and thus we also hope to use our analysis to help further this goal.

We used fMRI response data gathered for a set of image stimuli (1365 images from 8 different image categories) to build multi-layered perceptron (of different iterations) and logistic regression classifiers. These classifiers took in BOLD signal readings from individual voxels across the visual cortex and output a prediction for the image object category that the subject was viewing. We were able to achieve an accuracy of 44% from our MLP (chance = 12.5%) over 8 categories, which is substantially higher than any previous work that uses similar data/methods (Cox et al. 2003). For individual pairwise comparisons, we were able to achieve an accuracy of up to 96%. Finally, we were able to deliver insights about the relative importance of different regions of the visual cortex with regard to object category recognition and suggest avenues for further research based on our findings.

Related Work

Brain decoding is a relatively new field, spurred on by recent advancements in computer science and scanning technologies. The visual system is current focus of most studies, since it is relatively easy to carry out tests for understanding this area. In most experiments, participants are shown specific visual stimuli whilst in an MRI, and data is recorded from the visual cortex. One of the earliest papers in the field, by Kamitani and Tong (2005) focuses solely on decoding the orientations of gratings from the V1 region of the visual field. Our objective is more focused on decoding the content, rather than the orientation of stimuli, and also more complex in nature - we are concerned with natural images rather than simple gratings. The work remains however, of foundational importance in understanding the first stage of processing in the visual stream.

Kay et al (2008) demonstrate a method of identifying natural images from brain activity.They use a generative model to conduct voxel activity pattern prediction on a known set of images, and choose the image with the highest correlation to the actual voxel activity pattern recorded from the stimulus. Like us, they are concerned with decoding natural images, but their work is focused more on identifying the right image out of a set, rather than decoding the content of the image. Furthermore, we wish to accurately decode any natural image stimuli, not just those from a known prior set.

Cox et al (2003). attempt to classify object category from fMRI, but their study is far less rigorous than ours. They do not attempt to decode natural images (which is harder), and they do not test on more debated object categories, such as faces (which we do). Furthermore, they use far fewer training/test examples (100 vs our 1365), use a far smaller feature size (<100 voxels vs our 3k+) and their model is built using SVM pattern recognition. This results in far worse classification accuracy (<30%), than we achieve and a less applicable model, since in the real-world our brain is presented with natural images featuring objects at non-uniform angles (which we account for). Nonetheless it served as an important starting point to guide our approach.

Our work is built off experiments conducted by Seibert et al in “Modeling neural responses in the ventral stream”. Whilst we start with the same data, they focused on optimizing a deep convolutional neural network based on canonical computations to perform object recognition. Furthermore, their CNN is trained on images and the only reason they use fMRI data is to see whether CNN layers developed representations that corresponded to the different visual areas in the brain. Our approach however, focuses solely on prediction using fMRI data, because we are more concerned with building an optimal brain decoder.

Dataset & Features

Neuron responses to stimuli can be recorded via Functional Magnetic Response Imaging (fMRI). When neurons fire, they require more energy, which results in more oxygenated blood being direct to that area of the brain. This blood-oxygen-level dependency (BOLD) signal is what is measured with an fMRI. Because fMRI measures a secondary effect of neuron firing and lacks the same spatial resolution of measuring neuron activity directly with an electrode, fMRI can produce noisy data. This opens up the opportunity for the application of machine learning techniques.

Seibert et al. selected the two subjects which had the highest mean split-half reliability in V1 to complete a full data set (at least 9 sessions each consisting of approximately 10 8-minute scans of the main experiment). They presented 1785 gray-scale images of objects a median of six times across multiple sessions to each subject. Objects were drawn from 8 categories (animals, tables, boats, cars, chairs, fruits, planes, and faces) containing 8 exemplars. Each object was shown from 27 or 28 different viewpoints against a random natural background (circular vignette, radius 8◦ centered on fixation) to increase object recognition difficulty.

Our dataset is in the form of matrices, with each matrix representing a different region of the visual cortex. In each matrix, the columns correspond to each voxel in that region, and the rows correspond to each stimulus presented. In each of the two trials

1785 grayscale images across 8 categories were presented to each of the subjects. In our dataset we did some pre-processing by only keeping the images presented to both of the subjects, which left us to 1365 images. This also helped us to normalize the data across both subjects. To reduce noise, these filtered images were shown again to the same subject in a different trail such that each of these image would have been shown again in a separate trail. Our dataset was then comprised of stimuli responses to the shared image in both trials. This left us with a final stimuli of 2730.

Visual stream split into **12** regions - **V1, V2, V3, V3a, V4, FFA, LO1, LO2, LOC, PPA, OFA, TOS**.

Data Matrices: (stimuli x voxel)
V1: (2730 x 379)
V2: (2730 x 601)
V3: (2730 x 555)
V3a: (2730 x 148)
V4: (2730 x 350)
FFA: (2730 x 211)
LO1: (2730 x 118)
LO2: (2730 x 114)
LOC: (2730 x 460)
PPA: (2730 x 97)
OFA: (2730 x 125)
TOS: (2730 x 66)
Whole Visual Stream: (2730 x 3224)

When running our classifiers, we used 10-fold cross-validation analysis. This is something we hadn't done by the time of our poster presentation, but implemented after TA advice. We trained our models on 90% of the stimuli data, tested on 5% and validated on 5%. For each result, we ran our analysis 10 times, choosing a different 90% for the data for our training set and a different 5% of the data for our test set making sure we kept the same 5% of the data for our validation set. Our final accuracy score is average score from the 10 validation set accuracies.

Methods

For our model, we chose to use a multi-layer perceptron (MLP). A MLP is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. MLPs consist of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. These networks are good at learning complex nonlinear separating functions. MLPs train using back-propagation. More precisely, MLPs train using some form of gradient descent and the gradients are calculated using back-propagation. For classification, it minimizes the cross-entropy loss function, giving a vector of probability estimates p(y|x) per sample x.

We used the MLPClassifier class in the sklearn.neural_network library[2]. We utilized the Limited Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm, Stochastic Gradient Descent with Nesterov's Momentum[3], and Adam method for stochastic optimization[5] as solvers for our MLP, and found that using Adam as a solver produced the best results. Adam and SGD with Nesterov's Momentum both performed better than L-BFGS due to the fact that L-BFGS performs better on small datasets. We then tuned our hyperparameters using both cross-validation.

L-BFGS is an optimization algorithm in the family of quasi-Newton methods that approximates the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm using a limited amount of computer memory[4]. It uses an estimation to the inverse Hessian matrix to steer its search through variable space storing only a few vectors that represent the approximation implicitly. L-BFGS maintains a history of the past *m* updates of the position **x** and gradient $\nabla f(\mathbf{x})$. The updates are used to implicitly do operations requiring the $\mathbf{H}_k$-vector product.

The Adam algorithm updates exponential moving averages of the gradient and the squared gradient where the hyper-parameters control the exponential decay rates of these moving averages. The moving averages themselves are estimates of the 1st moment (the mean) and the 2nd raw moment (the uncentered variance) of the gradient. However, these moving averages are initialized as 0 vectors, leading to moment estimates that are biased towards zero. The initialization bias is easily counteracted with estimates of the gradient and the squared gradient. Adam's update rule is careful choice of stepsize, with two different upper bounds depending on the compute bias-corrected first moment estimate and the square root of the compute bias-corrected second raw moment estimate.

Stochastic Gradient Descent with Nesterov's Momentum is SGD using Nesterov accelerated gradient (NAG) which is a first-order optimization method to improve stability and convergence of regular gradient descent. It was shown that NAG could be computed by the following update rules:

$$v_t = \mu_{t-1} v_{t-1} - \epsilon_{t-1} \nabla f(\theta_{t-1} + \mu_{t-1} v_{t-1})$$
$$\theta_t = \theta_{t-1} + v_t$$

where θt are the model parameters, v_t the velocity, $\mu_t \in [0, 1]$ the momentum (decay) coefficient and $\epsilon_t > 0$ the learning rate at iteration t, $f(\theta)$ is the objective function and $\nabla f(\theta')$ is a shorthand notation for the gradient: $\partial f(\theta)/\partial\theta\,|_{\theta=\theta'}$

For our analysis we focused on answering the following questions:

1. How well does our model do overall when trained on fMRI data from all regions of the visual cortex on predicting image categories given fMRI data?

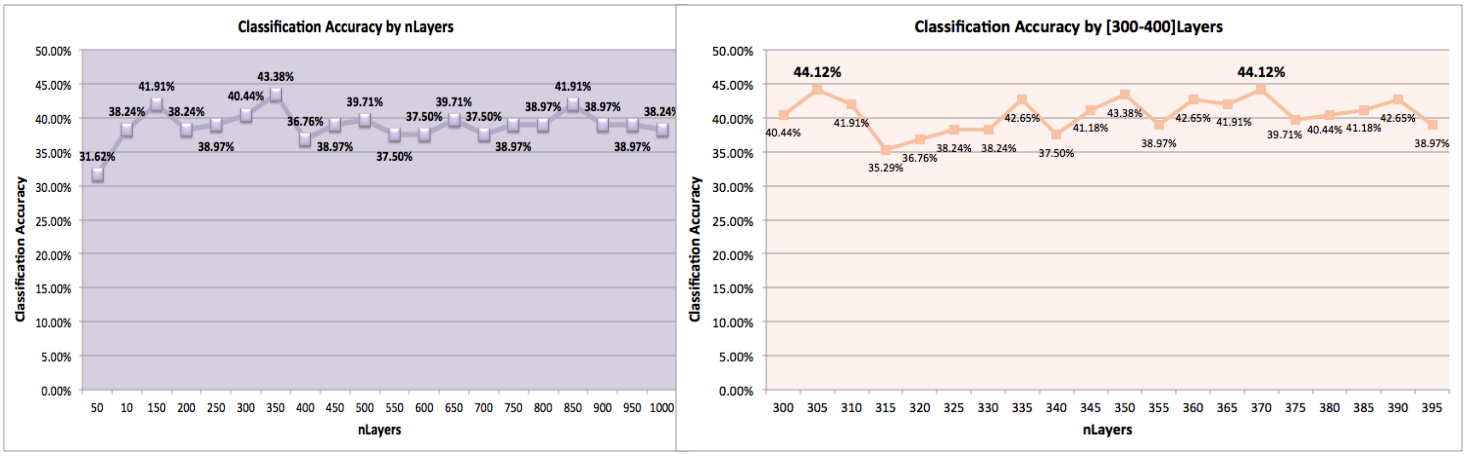
2. If trained on only two specific categories, how well does our model do at predicting those categories?
3. How does removing fMRI data from only one region affect the overall performance of our model?

We answered the first question by grouping together all the fMRI data from the different regions of the brain and separated them into a training set (90% train with 5% test) and validation set (5%). We then trained the optimized model on the training/test set and used the validation set to see the accuracy of this model.

We answered the second question by grouping together fMRI data from all regions of the brain, but only used 95% data from 2 distinct stimuli categories to train our model. We then tested on the remaining 5% (using cross-validation) to see how well the model did at predicting those 2 image categories. We did this for all possible combinations of pairs of the 8 categories.

We answered the final question by grouping together fMRI data from all regions of the brain except for 1. We then performed 10-fold cross-validation to find the optimized model and used the final average validation accuracy as the individual accuracies for each model. We did this for all possible combinations of 11 of the 12 regions of the brain.

Experiments/Results/Discussion

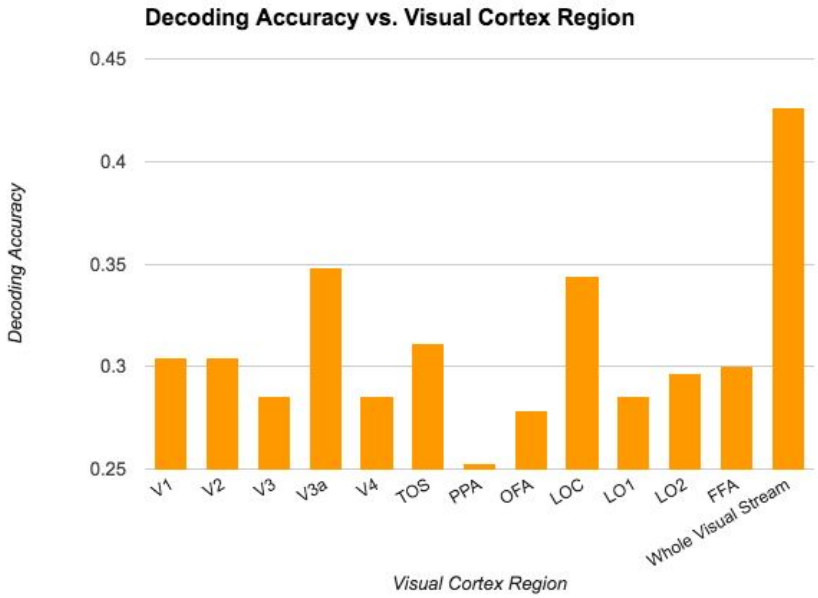


When optimizing our MLP Model we choose the following hyperparameters for Adam's method using 10-fold cross-validation:

- activation: rectified linear unit (ReLU)
- alpha: 1e-05
- beta1: 0.9
- beta2: 0.999
- epsilon: 1e-08
- initial learning rate: 0.001
- hidden layers: 305

We choose the optimal hidden layer amount by running our model (during the cross-validation analysis) on a series of layers from 50-1000 incremented by 50. We then found the highest accuracy to be 350. We ran the training a second time for layers between 300-400 and found the most optimal accuracy to be ~**44%** using 305 hidden layers.

In order to understand the impact of each visual region on object category recognition, we trained and ran our MLP classifier exclusively on each of our 12 regions. We wanted to know whether there was a specific region that specialized in object category recognition or if there is a region that makes no contribution at all. Furthermore, a good psychology rule of thumb is that further up the visual cortex you go (from V1 to FFA/TOS) the more complex (hence better) the visual processing. This is because we know from the Kamitani and Tong paper, that V1 mainly decodes orientation, whilst FFA can decode entire faces as a whole. The results for this comparison are shown below



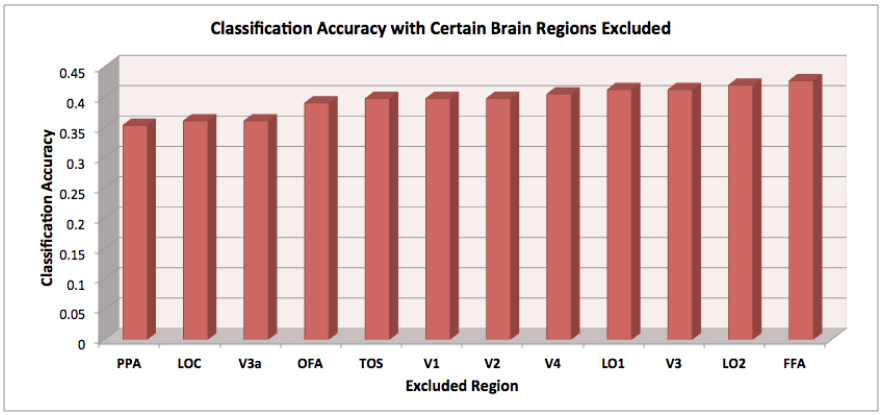
What we see to the left is that rather than just one region being solely responsible for object category recognition, all the regions have a relatively similar classification accuracy (range of 25% - 35%). The fact that decoding from the entire visual cortex results in an accuracy far greater than each individual region, suggests that

all visual regions have a part to play in object recognition. What is also surprising is that the earlier processing regions (V1-V4) have a similar, if not better decoding accuracy than the later visual regions (TOS - FFA). This could suggest, that orientation and other simple processing is more important for accurate object discrimination than more abstract levels of processing, or that earlier visual regions make a far greater marginal contribution in discriminating objects while later regions are far more dependent on earlier regions, meaning they are a bit handicapped on their own (e.g. LO1 + V1, is great, but LO1 by itself is really weak). It is interesting to note here that PPA had by far the lowest decoding accuracy of all regions.

To further shed some light on the relative importance of each visual region, we approached our analysis from a different angle. We trained and ran our classifier on the entire visual region except for one, changing out the excluded region on each run. This should tell us which brain region is the most crucial i.e the ones the others struggle the most without. The results of this are shown in the Classification Accuracy with Certain Brain Regions Excluded chart.

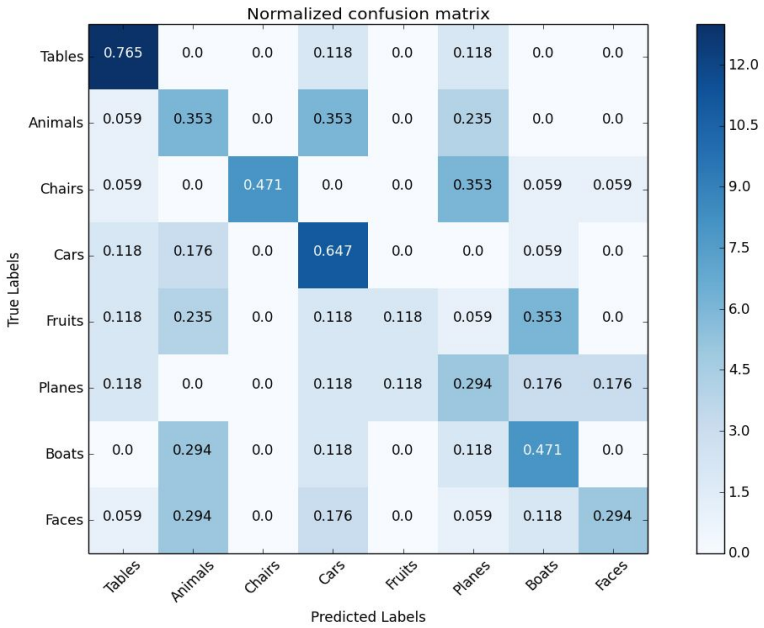
This shows a rather interesting result. Firstly, all the accuracies are largely the same (more so than the previous graph) which implies that no one region is paramount for object decoding. However, we can see that decoding accuracy fell the most (~10% drop) when PPA was excluded, but barely budged when FFA was excluded. Both these results seem rather counter-intuitive at first.

Focusing on the PPA, we saw earlier that it had the worst accuracy when decoding on its own (suggesting it is the least relevant region), but if we exclude it from our model based on the whole visual stream, our decoding accuracy falls the most (suggesting it is the most relevant region). Further analysis however, suggest this result makes perfect sense. The parahippocampal place area (PPA) plays an important role in the encoding and recognition of environmental scenes (rather than faces). Previous fMRI studies indicate that this region of the brain becomes highly active when human subjects view topographical scene stimuli such as images of landscapes, cityscapes, or rooms (i.e. images of "places"). We know that the image stimuli shown during this experiment were of objects imposed on a natural landscape background. Thus it must mean that the PPA isn't great at discriminating between different objects, but is essential for discriminating an object from the natural background scene it is in. Hence, without the PPA, it is much harder to distinguish between an object and the background, meaning that decoding accuracy falls overall.



Directly opposite to the PPA results, we can see that the LOC is one of the best individual regions in terms of decoding accuracy, and without the LOC, object categorization accuracy drops nearly as much as it does without the PPA. It is well known that the Lateral Occipital Cortex (LOC) is very important in object recognition, and it is good to see that our results do actually support this. Whilst we have shown that it is not solely responsible for object recognition, our results confirm that is a very important region.

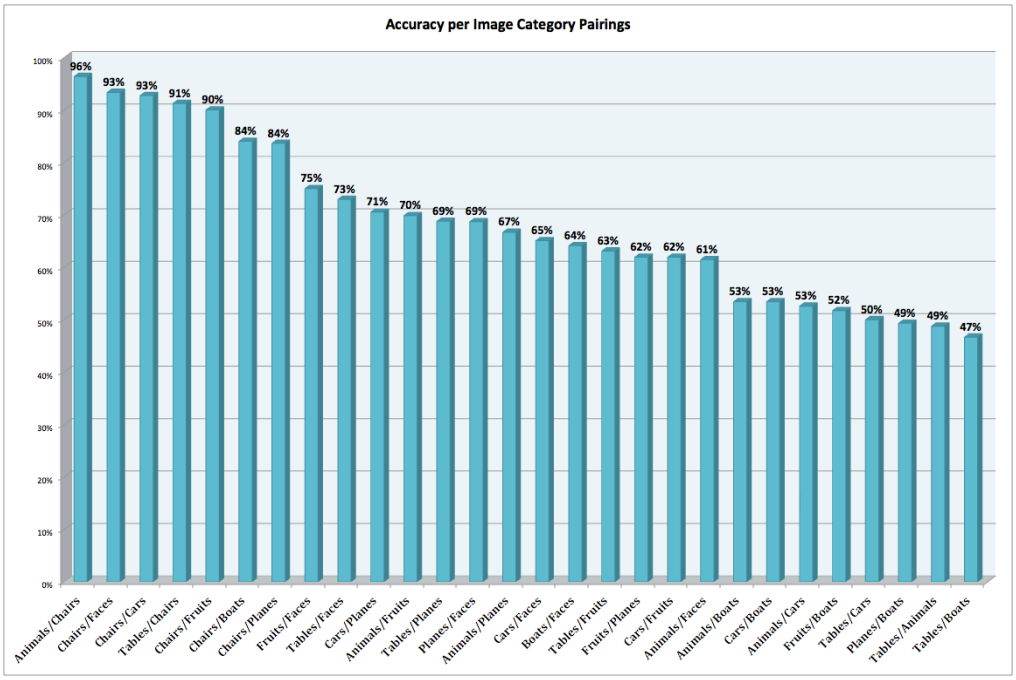
The fusiform face area (FFA) is known in psychology for being specifically specialized at detecting faces. Given that faces is one of our 8 categories, surely it is very important to our decoding accuracy? And yet this value barely drops when the FFA is excluded. Further research tells us that the FFA is not as useful in decoding faces from non-faces, but rather specialized in distinguishing between different faces. Hence someone with Prosopagnosia (a lesion to their FFA) can tell they are looking at a face, but not who's face it is. Hence at this level, we can see that our data confirms this theory. Future work to confirm this hypothesis would involve running our classifier just on the FFA to decode face vs non-face (should have moderate accuracy) versus running our classifier just on the FFA to discriminate between faces in our data set (should have high accuracy).



Stepping back a little, we wanted to know what was going wrong with our classifier. Though an accuracy of 44% is great, we would ideally want it to be 100%. Hence we built a confusion matrix (with a heat map) to see if our classifier was mis-classifying specific types of objects. In an ideal world, it would be a diagonal matrix.

Off the bat, we can see that our classifier is great at decoding some categories (such as tables or cars) but gets others completely wrong (such as fruits). And for some categories, they are easily confused for others. For instance, planes are easily confused for boats and cars, and faces are easily confused for animals. What does this all mean. We know that tables are made of perpendicular straight lines - which V1 is ideal for detecting. Hence by the time the entire visual stream has processed a table stimuli, we can be pretty sure whether we are looking at a

table or not. Perhaps more abstract object discriminations are handled outside of the visual cortex. To get a better picture of where exactly our classifier gets confused, we trained our classifier on the whole visual cortex, but discriminating only between all the possible object category pairs. The results can be seen in accuracy per image pairings.



Here again we can see that our model is poor at discriminating between animals and faces (61% accuracy vs chance of 50%) or even between cars and boats (53%) and even worse between planes and boats (49%). Given animals and faces are the only two categories of animate objects, perhaps our visual system merely categorizes them as such and later regions of semantic processing in the brain are left to distinguish what type of animate object they are. Likewise, cars, boats and planes are all methods of transport - indicating again that it could be the visual cortex merely classes them as "movement" or "transport" and later regions such as the temporal lobe are left to distinguish between them. A good way to test for this would be to run our classifier in the same way, but on the whole brain rather than just the visual cortex. If the decoding accuracy improves substantially, then it is highly likely that object identification is handled outside just the visual cortex. Alas, we did not have the whole-brain data available, and even then, the processing time for this would go through the roof. Nonetheless, this would be a great avenue for future research.

Conclusion

In conclusion, we were able to apply machine learning techniques to build a classifier to decode image object categories given fMRI data. We used a unique approach, a multi-layered perceptron, to classify images over 8 different object categories at an accuracy of 44% (chance = 12.5%). This is an improvement on previous fMRI object category decoding accuracy of < 30%, and on natural image backgrounds which are harder to decode. It is also important to consider that we are able to get this decoding accuracy even though we are measuring a secondary effect of neuron firing (blood flow in the brain) and are working at a spatial resolution of 1 million neurons. Direct neuron recordings from the visual cortex should give much cleaner, more reliable data.

We were also able to use our classifier in different ways to give insights into how our visual cortex works. We show that no one visual region is responsible for object recognition, but rather that they all work together. We also discount the idea that upper levels of the visual cortex (e.g PPA, FFA) are necessarily better at recognizing objects as a whole than lower regions (V1-V4). We show that the PPA, while not useful for object discrimination, is extremely useful for discriminating an object from a natural background - allowing the rest of the visual cortex to more easily categorize the objects. We also confirm the importance of the LOC for object recognition.

We believe our implementation of a multi-layered perceptron is a highly useful tool for both determining classification given brain data and carrying out analysis of different brain regions. We hope it is used in future brain decoding endeavors and are confident that, given a feature set of voxels from the entire brain, rather than just the visual cortex, we can achieve much higher classification accuracy, and prove whether object recognition takes places in brain regions further afield from the visual cortex.

Future Work

With more time, machine learning team members, and computation resources we would have done more hyperparameter optimizations other than k-fold cross-validation specifically using following Bergstra and Bengio's *Random Search for Hyperparameters Optimization*. We also would have worked further with Seibert et al. to see how we could develop some derived features for our model. We would perform more research on different variations of neural networks to ensure that the MLP Classifier best fit our practices. Finally, we would've worked with Seibert et al. to further filter out the noisy fMRI data.

We were interested in further pursuing our observations of how the FFA works, by running our classifier on a face vs not face object recognition task compared to a face A vs face B recognition task. This would shed more light on the extent of whether the FFA is solely focused on discriminating between faces, and help advance our understanding of conditions like Prosopagnosia. We would like to run our classifier on the entire brain to see what non-visual cortex regions are important to object recognition. We would also like to test our hypothesis for whether similar object categories (e.g. cars, boats, planes) are differentiated outside the visual cortex.

References

1. Seibert, Darren, Daniel Yamins, Diego Ardila1, Ha Hong, James J. DiCarlo, and Justin L. Gardner. "A Performance-optimized Model of Neural Responses across the Ventral Visual Stream." (2016): n. pag. Web. 13 Dec. 2016.
2. Scikit-learn Developers. "Multi-layer Perceptron." 1.17. Neural Network Models (supervised) — Scikit-learn 0.18.1 Documentation. N.p., n.d. Web. 13 Dec. 2016.
3. Kingma, Diederik, and Jimmy Ba. "A Method for Stochastic Optimization." Cornell University Library, 23 July 2015. Web. 13 Dec. 2016. <http://arxiv.org/abs/1412.6980>.
4. Nocedal, J. (1980). "Updating Quasi-Newton Matrices with Limited Storage". *Mathematics of Computation*. 35 (151): 773–782. doi:10.1090/S0025-5718-1980-0572855-7
5. Kingma, Diederik, and Jimmy Ba. "A Method for Stochastic Optimization." Cornell University Library, 23 July 2015. Web. 13 Dec. 2016. <http://arxiv.org/abs/1412.6980>
6. "CS231n Convolutional Neural Networks for Visual Recognition." CS231n Convolutional Neural Networks for Visual Recognition. Stanford University, n.d. Web. 13 Dec. 2016. <http://cs231n.github.io/neural-networks-3/>.
7. Cox. "Functional Magnetic Resonance Imaging (fMRI) “brain Reading”: Detecting and Classifying Distributed Patterns of FMRI Activity in Human Visual Cortex." *Functional Magnetic Resonance Imaging (fMRI) "brain Reading": Detecting and Classifying Distributed Patterns of FMRI Activity in Human Visual Cortex*. N.p., n.d. Web. 17 Dec. 2016. <http://www.sciencedirect.com/science/article/pii/S1053811903000491>.
8. Kamitani, Yukiyasu, and Frank Tong. "Decoding the Visual and Subjective Contents of the Human Brain." *Nature Neuroscience* 8.5 (2005): 679-85. Web.
9. Kay, Kendrick N., Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. "Identifying Natural Images from Human Brain Activity." *Nature* 452.7185 (2008): 352-55. Web.