How the Human Brain Understands Natural Scenes

Nathan Heck
ECE Department
University of Florida
Gainesville, FL
nathan.heck@ufl.edu

Jackson Cornell
ECE Department
University of Florida
Gainesville, FL
jcornell@ufl.edu

Justin Schubeck
ECE Department
University of Florida
Gainesville, FL
jschubeck@ufl.edu

I. Motivation

When human eyes view objects in the world, large swathes of neurons are activated in the outermost layer of the brain. Observing the brain activity of an individual while they are viewing different images could lead to learning information about how the brain interprets visual stimuli. The Computer Science and Artificial Intelligence Lab (CSAIL) at MIT is seeking to make progress in this field of research by issuing the 2023 Algonauts challenge. They have provided a dataset of fMRI signals that were recorded while presenting experimental subjects with visual stimuli from the COCO images dataset. Challengers are allowed to use any learning model they choose. The model must be trained on the given fMRI signals, and when presented with a test image as its input, will synthesize a set of signals that closely resemble the recorded brain activity as seen in Fig. 1. Entries into the challenge are judged by how the model's generated signals correlate with the fMRI recordings in the dataset across all points of interest in the brains of the 8 experimental subjects.

The driving factors behind this project are to foster innovation in the fields of biological engineering and machine learning to predict human visual brain data. The 2023 Algonauts challenge will lead to understanding the brain in new and better ways through state-of-the-art computational models for fMRI data. Our goal will be to provide the best model possible in this challenge to support the effort in advancing the relationship between biological systems and artificial intelligence.

II. Dataset

We will be working with the Natural Scenes Dataset (NSD) which consists of 73,000 different images and corresponding fMRI recordings. fMRI stands for Functional Magnetic Resonance Imaging and is a non-invasive procedure. fMRI has good spatial resolution in the brain compared to other non-invasive recording methods like electroencephalography (EEG). The hemoglobin molecule in our blood is responsible for carrying fresh oxygen to the parts of our brain that are being exerted. Hemoglobin has different magnetic properties depending on whether or not it is attached to oxygen. FMRI machines map out the volume of the brain into small, standardized units called voxels, and the magnetic sensors track how much oxygen carrying hemoglobin is present in each voxel. Each recording will usually contain measurements for close to 100,000 voxels, which correspond to about a 3 x 3 x 4 mm section of brain tissue.

This experiment is only concerned with certain groups of voxels along the visual cortex, termed Regions of Interest (ROI). The data given for the challenge comes from a large-scale experiment where fMRI recordings were taken on subjects while showing them images from NSD. The 8 participates were shown 9,000 unique images and 1,000 of the same images. Each image was shown 3 times for a total of 30,000 trials per subject. Each trial has a corresponding fMRI scan of the right and left hemisphere of the brain, represented by $\sim$19,004 and $\sim$20,544 voxels, respectively. The recording process for fMRI is fairly slow, so it would have likely required an extended period of time and effort for CSAIL to capture data for more than eight subjects. [1] [2]

The dataset is further split into training and test sets. For the training set, each of the 8 subjects have 9841, 9841, 9082, 8779, 9841, 9082, 9841, and 8779 images and fMRI scans, respectively. Likewise, the test set consists of 159, 159, 293, 395, 159, 293, 159, and 395 sets of images. The voxel data for the test set is withheld to be evaluated for the competition. [2]

The Pearson's correlation will be used to evaluate model performance on the validation set. The test set held by the competition uses a modified version of Pearson's correlation given below.

$$R = \frac{\sum_i (x - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_i (x_i - \mu_x)^2 (y_i - \mu_y)^2}} \tag{1}$$

$$metric_{test} = \text{median}\{\frac{R^2}{NC_1}, ..., \frac{R_v^2}{NC_v}\} \tag{2}$$

III. Questions

- Can an ensemble model method allow for better fMRI predictability with fewer training samples?
- Can an autoencoder trained on the image dataset provide a sufficient latent space to predict fMRI response?
- How do linear regression models vs. neural network regression models compare in predicting fMRI response?

IV. Methodology

To answer these questions, our team began with an exploratory research period where we covered literature that dealt with machine learning and fMRI. Techniques that were

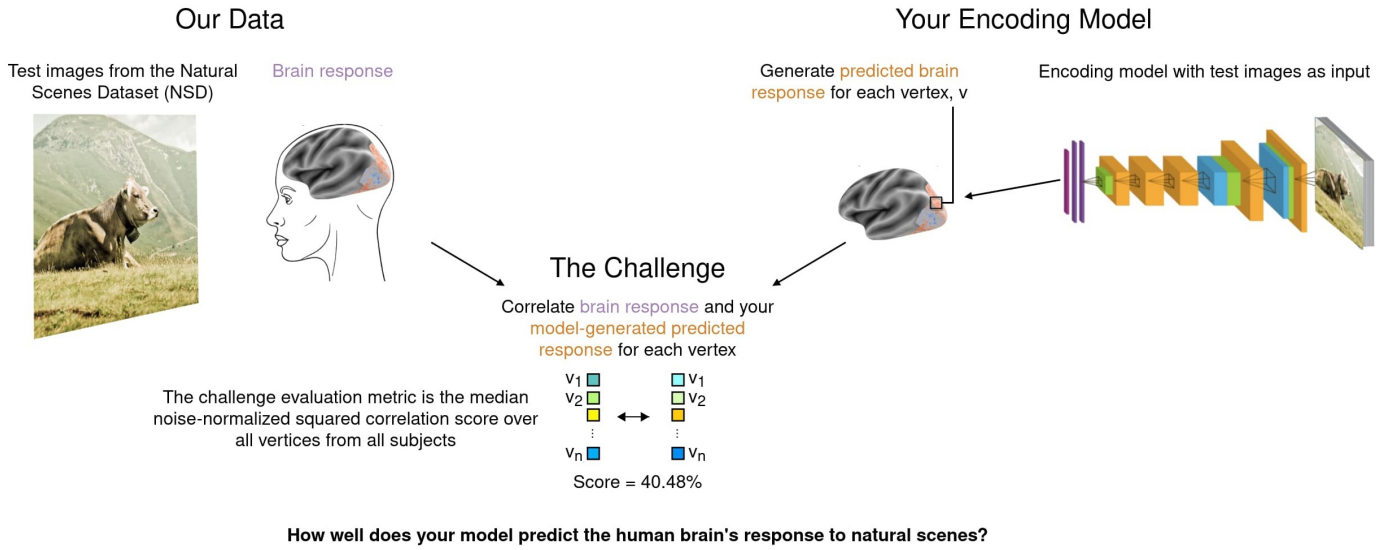

Fig. 1. Flow diagram of Algonauts image encoding challenge

used were noted. After several models were decided, our team ran the tutorial notebook given in the challenge to retrieve all the data and to understand how the organizers recommended interacting with the data in code. Our team then implemented our own learning model using similar techniques to those in the papers. We also decided to simultaneously implement a backup model using a different technique to compare to the first method's performance. A pipeline was established for training, validating, and testing our models with the given dataset that enabled us leverage our machine learning knowledge to make improvements to the model designs to improve performance.

V. Mixing Supervised and Unsupervised Learning

The first method we tried was mixing supervised and unsupervised learning. First, a variational autoencoder (VAE) was trained on a portion of the COCO image dataset, similar to Homework 2 where a VAE was trained on frames from the video of the mouse experiment. We used the VAE to encode each image into a latent variable space of different features that represent key points in the scene. The output of that encoder trained a regression model that predicted the activity in the brain by simulating the fMRI recording that would be generated if an individual was shown the same image. The encoded images were used rather than raw images to train the regressor because this performs dimensionality reduction on the input images, thereby lowering the training time required by the model and allowing it to hone in on the distinguishing features in the scene. Using this new representation of the images, both a linear regressor and an artificial neural network (ANN) were trained to predict the fMRI response at each voxel. The performance of these two methods were compared.

Initial processing was performed on the images using a function from the Algonauts tutorial code that resizes each image to a standard 224×224 (with three color channels), converts each image to a tensor, then normalizes each color channel which removes the light/shadow effects present across several similarly colored pixels. 90 percent of images in the COCO dataset were used to train the autoencoder, with the remaining 10 percent used to test the effectiveness of the feature extraction. This training code was run for each subject, and 2 trials each were performed to experiment with feature vectors, or latent spaces, of length 100 and 256. After the VAE passed the testing phase, it was used to encode all of the test images from the COCO dataset.

To train the regressor portion of the model, a train/test split provided by the competition of 90/10 was used. Then, an additional train/validation split of 80/20 was performed on the aforementioned training data. As the naming suggests, the regressors were trained on the training data, mapping the feature vectors to voxels of a given hemisphere of the brain. Both a linear and ANN regressor were trained for the right and left hemispheres of the brain for every subject. Validation was done by calculating the output of the validation set and applying the metric given in equations 1 & 2.

VI. Transfer Learning and Fine Tuning

The second approach to this mapping task was to use an existing model via transfer learning and fine tune it to the problem's application. In this case, there were many models to choose from: VGG16, VGG19, ResNet50, IncpetionV3, MobileNet, etc. For this methodology, ResNet-50 and Xception were fully implemented. The generic structure for transfer learning in this project can be seen in Fig. 2 with a ResNet-50 example for one individual.

A. ResNet-50

The preprocessing for ResNet-50 consisted of downsizing the images from (425, 425, 3) to (224, 224, 3), as the model expects this structure. The channels were also converted from RGB to BGR. Finally, the color channels were zero-centered

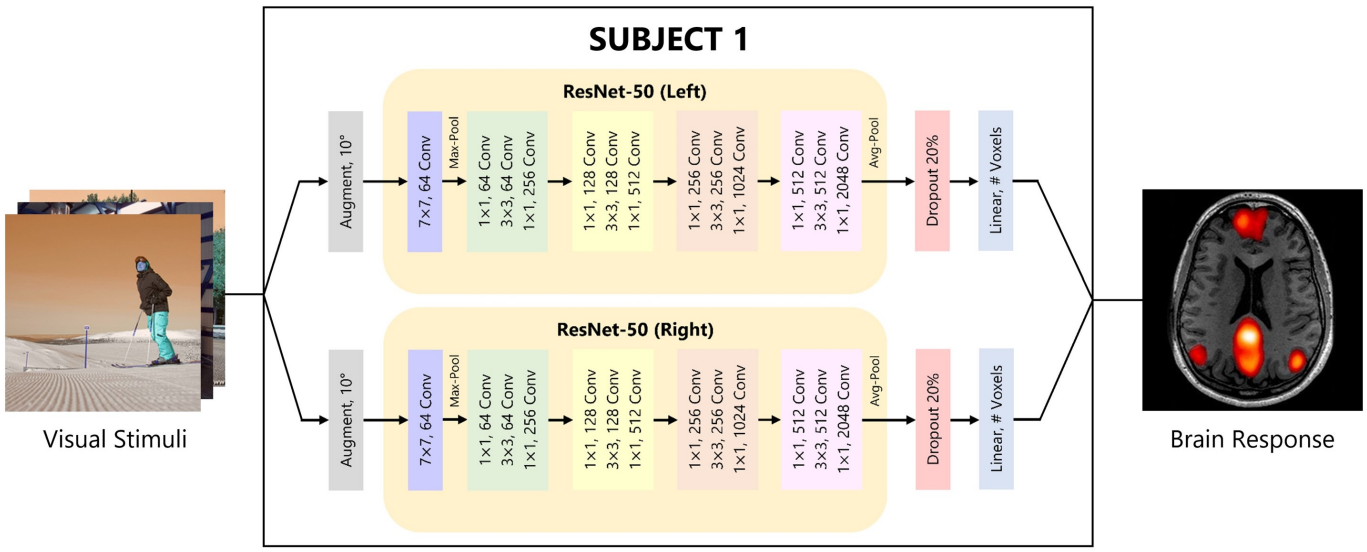

Fig. 2. Transfer learning structure (ResNet-50)

about the ImageNet dataset, which is what the ResNet-50 transfer learning weights were trained with. The data was also augmented by a 10-degree rotation and horizontal flip.

For the model's implementation, the base model was instantiated with the model frozen to keep the transfered information. The top layer was not included, and project-specific layers were added that included average pooling, dropout, and linear layer that represented the number of voxels. This structure was executed twice for each individual, representing the left and right hemisphere. The model was then trained on the training split for 20 epochs using the default Adam optimizer and MSE loss. Next, the model was unfrozen and retrained for 10 epochs with a learning rate 100x smaller to fine tune the model to this dataset. This training procedure can be seen in Fig. 3.

To address the proposal question of ensemble learning, each of the models trained for the subjects were ensemble averaged across voxels to test prediction correlation. This should theoretically generalize the brain, as it was testing across different individuals the responses to visual stimuli.

B. Xception

The preprocessing for Xception consisted of downsizing the images from (425, 425, 3) to (299, 299, 3), as the model expects this structure. The inputs pixel values were then scaled between -1 and 1, sample-wise. The data was also augmented by a 10-degree rotation and horizontal flip.

The procedure for training subject-specific models was the same as the ResNet-50 model. The Xception model by definition has more image resolution, higher top-1, and higher top-5 accuracy on the ImageNet dataset. Xception also has a smaller size, number of parameters, and depth than ResNet-50.

VII. Results

As a benchmark, we used a model consisting of AlexNet for feature extraction, PCA for extracting image features to a lower dimensionality, and linear regression for mapping the features to the voxels. These results can be seen in Fig. 4.

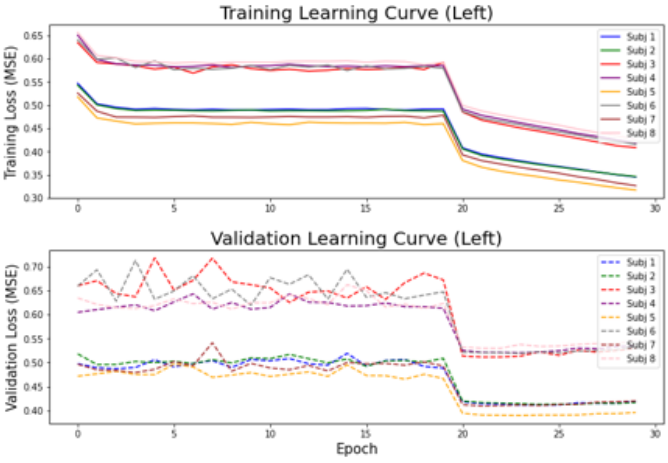

Fig. 3. ResNet-50 subject 1 left hemisphere training

A. VAE Model

Performance for the VAE model was underwhelming. In terms of the metric from equations 1 & 2, the VAE performed worse than the baseline model. Model performance was evaluated using both a linear and an ANN regressor trained and validated on subject 1. The better performer of the two, the ANN, can be seen in Fig. 5.

Besides its poor performance compared to the baseline, there are several observations one can make. An initial fear was that the ANN model may have overfitted. Plots of the training results alleviated this suspicion, as results were similar for training and validation. One may also note that, for the case of the right hemisphere, there was little difference in performance between the linear and ANN regressor validation

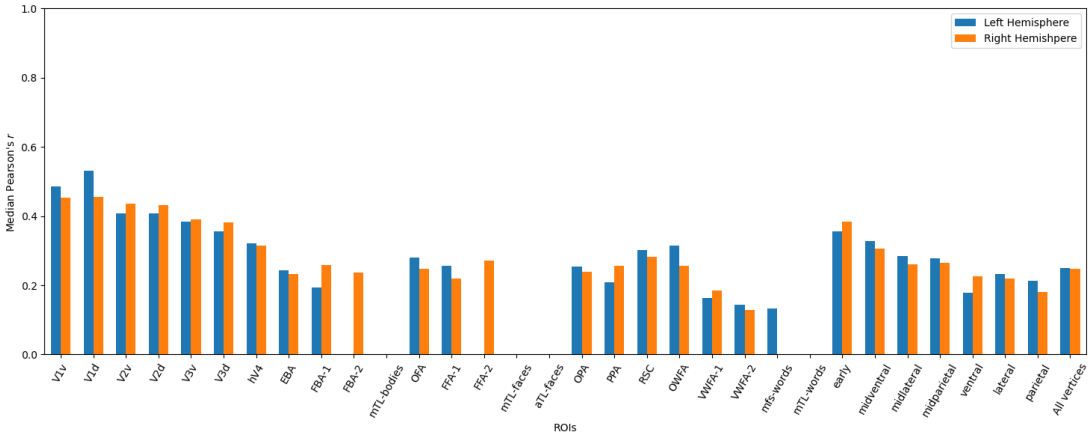

Fig. 4. Subject 1 baseline results on AlexNet with Linear Regression

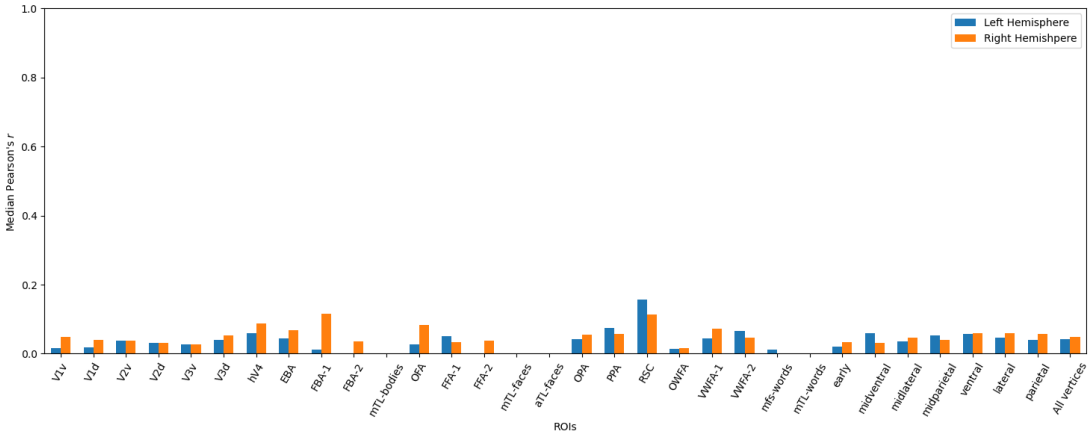

Fig. 5. Subject 1 validation results on VAE with ANN

results. This indicates that, unless it is completely necessary to squeeze out small performance gains, a linear regressor is ample for most applications of voxel prediction.

B. Transfer Learning Models

Overall, the performance of the transfer learning models was exceptional. The Xception model resulted in better performance than ResNet-50 when being applied overall to the 8 individuals. However, it could be possible that there is a specific architecture of transfer learning that would work best for each individual or each hemisphere. In terms of the challenge organizers' performance metric, ResNet-50 performed at 37.59 and Xception at 44.95. The Xception result placed our group 11th out of all submissions across the globe competing.

The second conclusion from the transfer learning model was that the ensemble averaging performed worse on average for each individual. The 6 individuals who had the same number of voxels in the dataset had their trained models averaged to create a more generic model. While this model may trade off some correlation performance, it may have generalizable capabilities for new individuals and their voxel responses. The comparison between the individual results and the ensemble results can be seen in Fig. 6.

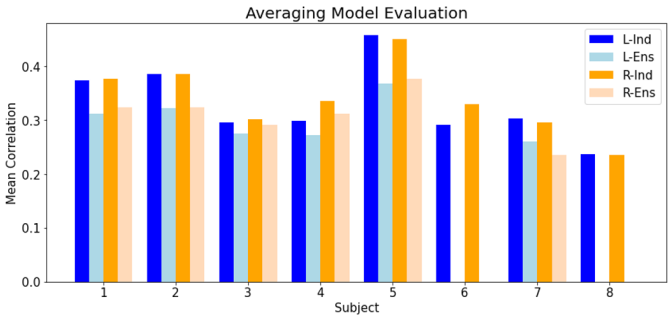

Fig. 6. ResNet-50 individual vs. ensemble performance

VIII. Obstacles

The biggest obstacles faced were computational; more specifically, training times and computational unit costs. For both the VAE and transfer learning models, computational units had to be purchased multiple times, which caused the

overall cost of this project to be higher than predicted. Additionally, both the time and computational resources needed to train these models made it difficult to iteratively change and re-test the models. This was especially challenging with the VAE model, as it was trained from scratch on the dataset.

IX. Division of Labor

All group members did our own research regarding fMRI and the data related to this challenge. We also conversed to pool information gained about any encoding models that would be relevant to this project. A final decision on the learning model types were made as a group, and these models were implemented as a team via a shared coding repository in Google Drive and Google Colab. Jackson took the lead on implementing the VAE plus regression models, and Justin took the lead on implementing the transfer learning models.

X. Conclusion

For the VAE, we can conclude that performance is linked to the latent dimension size. Unfortunately, we only got to test smaller dimension sizes of 100 & 256, but between the two we noticed substantial improvement. We can also conclude that the ANN regressor offers little improvement over the linear regressor with the added negative of being much more computationally expensive to train. A non-linear regression scheme is only suggested if the user wishes to squeeze as much performance as possible, though we caution the possibility of overfitting.

For transfer learning models, there are some structures that perform better at this task than others. The difference in the preprocessing of the images for ResNet-50 could be the reason it performed worse than Xception, as crucial feature information might have been lost. It also might also be a fact that a simpler, smaller model performs better in this case. Fine tuning the model was a crucial step as well, as the plateau of training error dropped significantly when adapting to the specific dataset.

Overall, we found that transfer learning was quicker, cheaper, and better than training a decoder from scratch. We also found that a linear regressor is more than ample to predict fMRI voxels than non-linear, given a sufficiently performative feature extractor. Finally, we found that ensemble models might generalize to new individuals, but did not increase the correlation compared to a specific individual's model.

XI. Future Work

In retrospect for this project, we would use transfer learning on the VAE decoder in future implementations. Too much computation was wasted on training this portion of the model, which could have been better spent on on the regressor portion of the model. Additionally, it's implementation and training were quite complex, with a custom training loop being written for it. We suspect using a larger latent space could also have provided better performance, as we saw a noticeable performance boost when increasing the latent space from 100 to 256. Again, we did not have time, or want to spend the money, to retrain the VAE with a large latent space.

For the transfer learning model, it would be interesting to compare more model types. We were limited to some model structures that use higher pixel spaces due to RAM capacity. Having more information from the image might lead to better feature extraction to predict voxel activity better. However, this could also lead to overfitting depending on the structure, as the ResNet-50 has more weights and depth than Xception, but Xception performed better. In terms of the competition, there is likely a model structure that works best for each subject/hemisphere pair. Fine tuning a model for each pair would take extensive hyperparamter tuning and model testing in order to reach the top score in the leaderboards, but would lead to interesting results going forward.

References

[1] Gifford AT, Lahner B, Saba-Sadiya S, Vilas MG, Lascelles A, Oliva A, Kay K, Roig G, Cichy RM. 2023. The Algonauts Project 2023 Challenge: How the Human Brain Makes Sense of Natural Scenes. arXiv preprint, arXiv:2301.03198. DOI: https://doi.org/10.48550/arXiv.2301.03198

[2] Allen EJ, St-Yves G, Wu Y, Breedlove JL, Prince JS, Dowdle LT, Nau M, Caron B, Pestilli F, Charest I, Hutchinson JB, Naselaris T, Kay K. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and computational intelligence. Nature Neuroscience, 25(1):116–126. DOI: https://doi.org/10.1038/s41593-021-00962-x

[3] Kingma, Diederik P., and Max Welling.m2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. DOI: https://arxiv.org/abs/1312.6114

[4] Gu Z, Jamison K, Sabuncu M, Kuceyeski A. 2022. Personalized visual encoding model construction with small data. ArXiv preprint, arXiv: 2202.02245. DOI: https://doi.org/10.1038/s42003-022-04347-z