Introduction: This can be copied from the proposal.

In this project, we are aiming to explore the correlation between layers of CNNs trained on psychophysics behavioral tasks and layers in the visual processing pathway through recorded electrophysiological (EP) data. This research question falls under interpretability of classification models. We were inspired by the paper, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," by DiCarlo et al. However, because this paper does not provide code or data, we are using the data from "Feature-selective responses in macaque visual cortex follow eye movements during natural vision," by Xiao et al.

Pre-existing models were either specialized for optimization of image classification or prediction of area IT activity. The hierarchical modular optimization (HMO) model has both task performance and area predictivity. Figure 1b demonstrates that models that are better at image classification are better representations of area IT. Specifically, as shown by Figure 2b, the HMO model is better at classifying images with high variation, such as different camera positions and lighting.

However, since the image dataset is not available in the paper, we aim to explore whether a similar model can be replicated with different data. In the "Feature-selective" paper, we found both task and EP data recorded from six areas in the macaque monkey ventral visual processing pathway. However, the data was collected from free-viewing tasks, so the data are trials of fixations and saccades of several thousand images. We decided to use the data on fixations and their classification as face versus non-face, such as in figure 2a.[1]

Challenges: What has been the hardest part of the project you've encountered so far?

The data has been the most difficult and frustrating part of the project. Especially since we had to diverge from the original paper due to the data availability, it has been difficult to figure out what data would be worthwhile to use. Moreover, using raw EP data and manipulating it into a form that is useful for validating it against the layers of the model is also difficult.

Insights: Are there any concrete results you can show at this point?

In order to test the model's performance on simpler data, we ran it on a simple binary cat-dog classification task from Homework 3 (ResNet data). In order to save on computation, and since we are mostly using this data to ensure that the model runs correctly, we did not train a large sample size of initial convolutional networks as our first layer. Instead, we just trained 10 that performed decently (~55-62% accuracy) on randomized hyperparameters like we will do when we train a full set. After putting these layers in the full HMO model and training it, we have received 65-75% accuracy on the simple classification task, varying per run.

While this is not an amazing score, we recognize that the point of the project is not to build the best image classifier. Our network architecture is optimized for interpretability comparisons with the human brain rather than image classification. After training the initial CNNs, the HMO model takes about 15 minutes for the 10,000 images to process.

How is your model performing compared with expectations?

[1] Further explorations could include using the saccade data and multi-label classification.

Right now, it is difficult to evaluate how “well” the model is performing since as mentioned, image classification accuracy is not the correct measure of model success for this project. We feel that as long as the model is able to classify images at a decent rate (for binary classification, >65%), which it does so far, we can be sure that the model is working. From here on out, we will be evaluating the model’s performance by comparing layer outputs to neural activations as done in the paper.

Plan: Are you on track with your project?

I would say that we are slightly behind since we wanted to have a fully trained model by this check in. However, since we are prepared to train the model, I wouldn’t say that we are too behind on our project.

What do you need to dedicate more time to?

We need to train our model, and then figure out how to extract the layers of the model to compare it to the EP data.

What are you thinking of changing, if anything?

We changed our plan of using fixations to just creating a classification model with more variety of data. Figuring out how to match the fixation data to meaningful labels did not seem to be a good way to spend our time. So, we are using image data like pareidolia images to try to get the high variation images of the original paper that we had originally tried to do with the fixations.