**Introduction**

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]

**Related Work**

The paper "Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition" by Cadieu et al. is a similar research paper to the one that we are reimplementing, with a few key differences. This paper is more focused on the IT cortex as opposed to including the V1 and V2 cortexes as well. Additionally, this paper uses a different accuracy metric than the Yamins paper. This paper simply uses classification accuracy compared to the classification accuracy of the brain's layer, which is a bit more simplified than the regression-based approach in Yamins. The key difference for our purposes between these papers is that Cadieu et al. simply train their classifier on AlexNet/ResNet images and use a more standard convolutional network approach for training. This would be a good alternative method for us if training the HMO model is impractical.

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003963

**Methodology**

The architecture of the HMO model in DiCarlo et al. are two stacks of three-layer CNNs that have been combined in parallel, with a 1250 top-layer output. While the basic architecture will not be difficult to implement, it will be more difficult to combine each layer of the individual CNN outputs, and then stack them. In particular, combining and reshaping the inputs will take some considerable work. Moreover, the original HMO model uses boosting and hyperparameter optimization, which will be considerably difficult and perhaps

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

not possible within the scope of this project. Thus, one “reach goal” would be to implement at least one of these training methods.

**Metrics**

After training the model on classification of images, we plan to examine the interpretability of it — not necessarily the accuracy of our model at the classification task. Instead of measuring accuracy, we plan on determining how well the model replicates the human/monkey brain on similar classification tasks. After each layer of the HMO (corresponding to V1/V2, V4,  and IT respectively), we take the outputs of the layer, flatten them, and use a regularized linear regression (ridge penalty) with the outputs as our independent variable and neural response (from the data) as our dependent variable.

From here, we can run multiple experiments. The primary one that the Yamins paper deals with is the Pearson correlation coefficient, r, which seeks to determine how related the neural responses are to our layer outputs. Of course, we can also use r^2 to determine the level of variance in neural response explained by our model outputs at each layer.

At this stage in the project, it is hard to assess our base, target, and stretch goals, but for now, we will aim for achieving a similar amount of correlation that the Yamins paper did (of course without overfitting to the neural responses).

**Data**

We are using the DANDI dataset from the study "Feature-selective responses in macaque visual cortex follow eye movements during natural vision", available via DANDI and documented in the associated GitHub repository. This dataset includes neural recordings from macaque monkeys viewing natural images, along with corresponding eye-tracking and TTL event data. The visual stimuli are packaged as multi-part zip archives (Stimuli.zip and Stimuli.z01) due to OSF file size restrictions. These can be extracted using tools such as 7z or PeaZip to access the original image stimuli.

Significant preprocessing is required to align the data temporally and spatially—specifically, to determine what image the monkey is seeing at each time point, and to identify which regions of the brain are activated in response. This involves parsing event timing, matching eye-tracking data to visual stimuli, and interpreting neural signals from different brain areas. Once aligned, we use this data to train a convolutional neural network (CNN) and apply interpretability tools like LIME to compare the network’s internal feature activation with the monkeys’ observed neural responses. Our aim is to explore the similarity between artificial and biological visual feature recognition under natural viewing conditions.

**SRC**

Our dataset involves neural recordings from macaque monkeys as they viewed natural images, with data collected under approved animal research protocols. However, the use of non-human primates in neuroscience raises valid ethical concerns. Macaques are intelligent, social animals, and procedures involving neural implants

and behavioral conditioning can cause stress or discomfort, even when performed under institutional guidelines. While the dataset documentation doesn't indicate direct harm, the invasive nature of the recordings highlights the broader tension between scientific progress and animal welfare. As researchers, we recognize these tradeoffs and aim to approach our analysis with transparency and respect for the animals involved. Ideally, our work can contribute to future models that reduce the need for such invasive studies.

A relatively philosophical take on our project relates to the growing convergence between artificial and biological cognition. As AI systems like convolutional neural networks begin to approximate not only human-level performance but also the internal representational patterns seen in primate brains, it raises deep questions about agency, autonomy, and moral consideration. If a model trained on naturalistic stimuli starts to "see" in ways that are neurally analogous to a living brain—especially one as cognitively advanced as a macaque—at what point do we begin to attribute some form of cognitive status to these systems? While current AI lacks sentience or subjective experience, philosophical traditions such as functionalism argue that what matters is the structure and function of cognition, not its substrate. This invites speculation about the future of rights: if artificial systems come to replicate the informational and behavioral signatures of biological agents, might we someday owe them ethical consideration—not unlike the questions we already face in animal research? While far from a settled issue, these questions force us to think critically about the kinds of intelligence we're building, and the responsibilities that come with it.