Decoding the Cortex

Noah Syrkis

November 21, 2023

Abstract

Understanding how the brain encodes visual information is a key challenge in neuroscience. In this project, we attempt to address this challenge by constructing a multimodal encoding model based on the Algonauts Project 2023 dataset. In addition to the dataset's image modality, we incorporate a semantic feature vector that describes object categories contained in the image shown to the subject during the functional Magnetic Resonance Imaging (fMRI) data collection. We combine various linear modules to construct two models: one predicting the fMRI data from both the associated image and the image's associated semantic feature vector; the other predicting both the fMRI data and the semantic vector from the image alone. Bayesian hyperparameter optimization suggests that the latter approach could potentially enhance model performance during inference without increasing the number of parameters. The model's performance was evaluated using a 5-fold cross-validation strategy and the median Pearson correlation coefficient as the metric. The code for this project is accessible at github.com/syrkis/neuroscope.

Introduction

Vision is the principal modality through which we interpret and interact with our environment. It is a cornerstone of human experience, profoundly influencing not only our perception of the external world but also the rich tapestry of our inner lives. The

visual system is a conduit through which a significant portion of our cognitive processing occurs, shaping our understanding in a multitude of domains, often surpassing the influence of language in its scope and depth (McCarthy 2017).

This report aims to dissect the intricate relationship between vision and cognition and to illuminate how this relationship is being reinterpreted and reshaped in the age of artificial intelligence (AI). We will explore how advancements in AI and machine learning, particularly in the field of neural networks, are not only mimicking human visual processing but also enabling machines to 'see' and 'understand' in ways that were once the sole province of biological entities. Through this exploration, we seek to unravel how these technological advancements are redefining our understanding of intelligence, perception, and the very nature of human-machine interaction.

Visual processing is the principal modality through which we interact and decipher our environment. Over the years, substantial progress has been made in understanding how the brain processes visual information, with even surprising parallels observed between artificial and biological vision processing [Cite]. However, as reality can exhibit extraordinarily different visual fingerprints—from simple geometric shapes to complex landscapes and visual noise—any system capable of visual perception is necessarily complicated. Fully capturing this complexity and intricacy remains a challenge. It is this challenge that is the focus of the 2023 Algonauts Project[1]. The Algonauts Project's 2023 dataset is based on the Natural Scenes Dataset (NSD), which couples images from the Common Objects in Context (COCO) dataset (Lin et al. 2015) with fMRI responses to those images from various participants.

Neuroimaging techniques like fMRI have facilitated valuable insights into the neural correlates of visual perception. However, the potential of these techniques has been somewhat constrained by computational model limitations and the expense and time required to collect large-scale fMRI datasets. Amid these challenges, deep learning has proven to be a powerful tool, that has facilitated a better understanding and emulation of

[1] http://algonauts.csail.mit.edu/

human visual perception. Recent efforts to incorporate multimodality into deep learning models have opened promising avenues to bridge the gap between computational models and the brain's complexity.

The experiment presented here explores how an additional modality might contribute to developing a model of the brain's visual encoding system, *without* a large increase in complexity/parameter count. The additional modality used here is a semantic feature vector, derived from the COCO dataset, describing the object categories contained in each image. The two models we developed are tasked with 1) predicting the brain response given the image and knowledge of what is in the image, and 2) predicting the brain response and the semantic contents of the image. The second model is there as an additional source of potential confirmation of the usefulness of adding a second modality. The first model is the heart of our experiment, allowing us to explore the question: "*How does the inclusion of an additional modality have on the performance of a brain encoding model during inference?*" It is to attempt an answer to this question that we have constructed the two models and their corresponding unimodal baselines.

Background

Decoding of the visual cortex was first done by Haxby et al. (2001).

Visual information processing Visual information processing, characterized by its hierarchical nature and intricate interconnectivity, plays a vital role in our understanding of the brain and perception. Traditionally, the process is categorized into low, mid, and high-level processing, focusing respectively on elementary visual features, their conjunctions, and abstract representations ((**gonzalez-casillas2018?**),(**groen2017?**)). However, the hierarchical categorization is insufficient to capture the full complexity of real-world scene perception. It underrepresents the multimodal and interconnected nature of visual perception, particularly when processing complex stimuli such as natural scenes (Allen et al. (2022), (**groen2017?**)).

The development of fMRI has opened up the possibility to examine and visualize brain activity associated with visual perception in real-time(Allen et al. (2022), Haxby et al. (2001)). However, despite these advances, the understanding of the intricate interconnectivity in visual perception, particularly for natural scenes, remains limited. Furthermore, it is increasingly evident that a holistic understanding of visual perception requires the acquisition and analysis of massive amounts of data ((**chang2019?**), Allen et al. (2022)). To meet this need for data-intensive analysis in the field of visual neuroscience, deep learning has emerged as a promising tool.

Deep learning through visual encoding models The power of deep learning models in the field of neuroscience is attributed to their ability to process and learn about high volumes of data, their inherent flexibility, and their structure, which is inspired by mirroring the brain's hierarchical organization (((**kriegeskorte2015?**)); ((**kell2018?**))). In this field, deep neural networks (DNNs), particularly convolutional neural networks (CNNs), have been extensively used to predict brain activity in response to visual stimuli, known as visual encoding models. Findings from several studies demonstrate that DNNs and CNNs can accurately predict neural responses to various visual stimuli ((khaligh-razavi2014), (yamins2014), (cichy2016)). For instance, Zheng et al. (2021) applied a convolutional recurrent neural network (CRNN) to decipher the computational elements of the retinal circuit involved in interpreting natural scenes. Findings showed that recurrent spatiotemporal receptive fields of ganglion cells were key in encoding dynamic visual scenes and that the inherent recurrence of the model enhanced the prediction of the neural response ((**zheng2021?**)).

Despite the richness of information available in fMRI data, remarkably few studies have exploited deep learning encoding models to predict associated brain responses. One study, by Güçlü and van Gerven (2015), utilized a DNN trained to predict object categories of over a million natural images and then used it to predict fMRI BOLD responses to complex naturalistic stimuli. The study identified a gradient of increasing complexity in the ventral visual pathway and found that the receptive fields in this

region are attuned to object categorization (güçlü2015). In another study, Zhang et al (2019), constructed a new visual encoding model that was based on a unique combination of techniques: it employed transfer learning to use a pre-trained Deep Neural Network (AlexNet) and used a non-linear mapping to translate visual features into brain activity. The researchers found that their model yielded significant predictions for over 20% of the voxels in the early visual area and resulted in the outperformance of conventional linear mapping models, offering a new approach to leverage pre-trained visual features in brain activity prediction ((**zhang2019?**)). Notably, these DNNbased models can even outperform traditional hand-engineered models ((guclu2015), (kell2018), ((**zhang2019?**))), signifying a substantial leap forward in uncovering the complexity of visual information processing.

Multi modality in visual perception and deep learning Most current models rely on single-modal input data, typically the visual stimuli themselves. Yet, the evermore revealed complexity of visual perception indicates a need for a shift towards multimodal deep learning models to encode brain activity.

Multimodality, in the context of neuroscience and cognitive science, refers to the brain's ability to integrate and process information through multiple sensory systems by which humans perceive and interact with the world ((parcalabescu2021); ((**gibbons2012?**))). In the context of machine learning, multimodality carries a similar theme but is applied to data. It refers to utilizing and integrating multiple information sources to enhance algorithms' performance. Several studies have pointed out that multimodal learning can help build more robust models that provide richer information about underlying data patterns and create more complex feature representations (((**ngiam2011?**)); ((**gu2017?**))). As visual information processing involves a complex interplay of interconnected factors and systems ((**groen2017?**)), infusing deep learning models with a multimodal approach could provide a more comprehensive and rich perspective. Few studies have investigated the potential of multimodal deep learning for predicting brain activity. One study by Horikawa and Kamitani (2017) combined deep neural networks (DNN) and fMRI to

predict and decode perceived and imagined objects. The model incorporated both visual and semantic (textual) data to create a multimodal encoding model.Very few studies have done this with fMRI responses. To the best of our knowledge, this is a niche approach. This approach could provide a more comprehensive understanding of visual perception and its underlying neural correlates, particularly when dealing with complex stimuli such as natural scenes. Multimodality is also an increasingly popular topic, partylu due to the release of GPT-4 which includes image and text modality.

To summarise; there are various approaches to brain encoding, using deep learning. Some architectures are rather advanced, utilizing recurrent and convolutional neural networks. We have opted for using a simpler model as a building block in a modular, multimodal approach, to better gauge the unique contributions of a second modality, and simple experimentation.

Methodology

Our methodology is that of a supervised machine learning experiment. We have access to preprocessed fMRI scans showcasing the blood oxygen level-dependent (BOLD) response to a variety of images. Our primary objective is to construct a multimodal model that has as many parameters during inference as its unimodal counterpart, and yet better predicts the brain's response to a given image. This section outlines the steps and components involved in the execution of our experiment. It should be noted that, following the Algonauts Project, the median Pearson correlation between voxels in the ground truth and the prediction is used as the target metric (though not as a loss function).

Data The data underpinning our experiment is provided by the Algonauts Project (Gifford et al. 2023) and is initially derived from the Natural Scenes Dataset (NSD) (Allen et al. 2022). The NSD is currently the largest dataset of its kind, encompassing cortical surface vertices from the left and right hemispheres of eight participants' brains.

These vertices correspond to the neurological responses triggered by 73,000 COCO images used by the NSD, each image depicting natural scenes. In addition to category information for each image, the COCO dataset provides other valuable metadata such as object location boxes and caption lists. Our experiment focused on the COCO object category information. The images in the NSD contain 80 different kinds of objects, with most images containing multiple object kinds (for example a horse and a person). As per the Algonauts guide[2], we represented each image using the dimensionality reduction method principal component analysis (PCA) of the all the image's activations in the 2012 image model Alexnet's second layer (Krizhevsky, Sutskever, and Hinton 2012). As in the Algonauts guide, PCA was performed reducing each image to a vector of size 100.

Over a year, each participant in the NSD study was exposed to 10,000 unique images, with each image presented three times, resulting in 30,000 image trials per participant. The corresponding fMRI data comprises 19,004 and 20,544 voxels for the left and right hemispheres, respectively. These voxel counts were selected based on preprocessed, high-quality 7T fMRI responses measuring as BOLD response amplitudes. Also included in the dataset are region of interest (ROI) masks for each subject, which aid in extracting specific fMRI data from certain locations in the brain. The fMRI data has been mapped to Harvard's FsAverage atlas such that the voxels are comparable across individuals. We eliminated subjects 6 and 8 from the experiment due to missing data (voxel counts differed from 19,004 and 20,544 for the left and right hemispheres respectively). We thus trained on six subjects.

Models The purpose of our models is to infer the BOLD response from a given image. The architecture of our primary model involves taking a vector representation of an image x, and outputting a tuple consisting of the left hemisphere BOLD response y_{lh}, right hemisphere BOLD response y_{rh}, and a semantic feature vector y_c for optimization against the COCO data. This model is partitioned into four submodules, each an MLP processing one of the four variables (x, y_{lh}, y_{rh}, and y_c). Our baseline will be a unimodal

[2]https://colab.research.google.com/drive/1bLJGP3bAo_hAOwZPHpiSHKlt97X9xsUw

version of this model. We aim to test if including the semantic modality improves performance.

The first module, referred to as the image encoding module, maps the input image vector x onto a latent space, thereby generating a latent vector z, which is subsequently fed into the remaining three modules responsible for predicting the outputs. As suggested by the Algonauts challenge baseline, the latent vector z maintains a dimensionality of 100. Given that each hemisphere's voxel count is approximately 20k, the linear mapping from the latent space to the voxel space demands around 2 million parameters. Therefore, even with such a compact latent space, the minimum required parameter count is approximately 4 million.

Our second model used y_c as input, concatenating it with x. The purpose of this model was to gauge the potential of multimodality on the input side of the network. This model is not our main focus, but rather a test to gauge the usefulness of this particular kind of multimodality.

All hidden layers used the tanh activation function, dropout of 0.1, and weight decay of 0.0001 with the AdamW optimizer from Optax. The models were implemented in Jax with Haiku(CITE). The shared (first) module had two layers, with 100 units each, to create some flexibility as the input to all other modules (the latent vector z) flowed through that initial module. The rest of the modules mapped the latent vector input to whatever output dimension their modality had. The learning rate was 0.001 and the batch size was 32. Hyperparameter optimization was not done on the aforementioned hyperparameters due to computational constraints.

The primary model (with the auxiliary task of predicting y_c during training), had two experiment-specific hyperparameters, α and β, weighing y_c and whatever hemisphere was not being optimized for respectively in the loss function. The model used mean squared error for optimizing the fMRI predictions and binary soft f1 loss for y_c due to a heavy imbalance between categories. Using regular binary cross entropy would yield a

low loss by guessing all zeros, as most images contain only a few categories.

Incorporating category vector modality and semantic vector representation To unlock the potential utility of the semantic vector, we designed our experiment with a multimodal approach. This involved integrating the category vector modality (model 2) by concatenating it with the image vector derived from AlexNet, an auxiliary task to predict the category during training (model 1), and tuning the α and β parameters weighting the importance of the auxiliary tasks in the loss function. Additional motivation for the inclusion of the auxiliary modalities is the potential avoidance of overfitting; finding inappropriate shortcuts in the data becomes more difficult if the shortcuts also have to make sense of the semantic vector.

Model training, auxiliary tasks, and hemisphere balancing Two key hyperparameters, α, and β, were used to balance the different aspects of our model's performance. α controlled the balance between fMRI loss and category prediction loss, thereby providing weight to the auxiliary task of category prediction. This strategy was based on our hypothesis that having the model solve an auxiliary classification problem could lead to more generalized and versatile representations beneficial for the primary task of predicting fMRI responses. β modulated the balance between the losses of the two hemispheres. By tuning this parameter, we hoped to find out if there is a balance that might contribute to a better overall model performance on the subjects.

Hyperparameter optimization and loss function design The cornerstone of our experiment involves hyperparameter optimization, carried out using the Weights & Biases (wandb) sweeps with wandb's Bayesian optimization techniques. The loss function is expressed as $(1 - \alpha)((1 - \beta)Lossy_{lh} + \beta\ Lossy_{rh}) + \alpha\ Loss_{y_c}$, when optimizing for y_{lh} and flipping the β when optimizing for y_{rh}. α serves as a weighting factor determining the trade-off between the fMRI prediction task and the category prediction task, while β controls the balance between the losses of the two hemispheres.

Bayesian optimization and cross-validation To search for the optimal values of α and *beta*, we initiated a wandb sweep with Bayesian optimization and optimized concerning validation of left hemisphere correlation in one sweep, and validation of right hemisphere correlation in the other sweep. This strategy enables a directed search in hyperparameter space, making it a more efficient and effective approach for hyperparameter tuning than random search or grid search. Additionally, we employed a K-fold cross-validation technique for model evaluation, providing a more robust estimate of the model's performance and optimal hyperparameters. K was set to 5. Every fold for every subject ran twice to get samples during the Bayesian optimization.

Results and analysis

The following will showcase our numerical results, and do an analysis and brief discussion of these. The results seem to indicate a positive answer to our question about the potentially regularising effect of including an additional modality during inference. As **table 2** shows, focusing on model 2, the semantic vector does contain some useful information as including it seems to increase the mean median correlation between the voxels by about more than 0.1 during training compared to the unimodal counterpart of model 2. For model 2 the benefit during training of the additional input modality still seems to be positive, though it is an extremely small increase in mean median (0.005 and 0.0042 for left and right hemispheres respectively). Bear in mind that the median is computed across the entirety of the voxel vectors, which are 19,004 and 20,544 long respectively, so a subtle increase in the median might be significant. It might also *not* be significant necessitating further exploration.

Exploring our primary model (with potential for the auxiliary task during training) in **table 1** we see the mean median voxel correlations for the two hemispheres trained with and without α and β set to 0 (setting these parameters to zero turns the model into our unimodal baseline). The none baseline has $\alpha = 0.05$ and $\beta = 0.25$. **Table 1** shows us that our multimodal model outperforms the baseline on the test data, yielding

a median correlation that is about 0.1 higher across both hemispheres in both train and test data, except the test data on the left hemisphere that is only 0.07 higher in median correlation. Thus, it appears that all else being equal, there is more benefit to using the second modality as an auxiliary task during training, than to receiving that information during inference. To the extent that this finding is significant, we find it surprising, as direct access to the second modality during inference seems more information-carrying, than indirect during training. Further analysis would however be needed to explore the significance hereof. We see slight overfitting on both the baseline and the multimodal model 2, with slightly more overfitting on the baseline, again being an indication of a slightly regularising effect from the second modality.

An understanding of the hyperparameter search for α and β can be gathered from inspecting **Appendix E** where we see that the correlation between α and the median-based metrics is small, though positive. The correlation between β and the median metrics is also slight, but negative. However, as seen in **Appendix C** and **Appendix D**, the Bayesian hyperparameter optimization left the low β space slightly underexplored. Setting α and β values were made with human intuition having inspected the sweep logs. Again more elaborate statical analysis would be needed.

Table 1: Mean Median Voxel Correlation (Model 1).

Hemisphere	Train, Alex/COCO	Train, Alex	Test, Alex/COCO	Test, Alex
Left	0.2558	*0.2676*	*0.1869*	0.1812
Right	0.255	*0.265*	*0.1881*	0.1782

In **table 2** we see mean median voxel correlations across all subjects and folds of model 2 with and without the COCO vector concatenated to the Alex vector. We see a similarly subtle advantage to including the second modality here. The reader should again note that the metrics here displayed are mean *median* correlations.

Table 2: Mean Median Voxel Correlation (Model 2).

Hemisphere	Train, Alex/COCO	Train, Alex	Test, Alex/COCO	Test, Alex
Left	*0.2176*	0.2059	*0.1932*	0.1927
Right	*0.2155*	0.2046	*0.195*	0.1908

Also relevant is that the correlation between prediction and truth is not uniformly spread throughout the visual cortex, but rather concentrated on earlier regions of interest, as showcased in **Appendix A**, **Appendix B** and interactively at neuroscope.streamlit.app

Future Work

As seen in the previous section, it appears that the semantic vector modality is not particularly useful for the model. A logical next step would be to experiment with extracting the image representations from different, or multiple AlexNet layers, or using an entirely different model for the image representation extraction. We might also explore using more rich COCO modalities such as image captions and object bounding boxes. Lastly, from a neuroscientific perspective, the ROIs of the brain are considered to be different modalities: they function by vastly different rules. Processing the ROIs separately might allow for models tailoring to specific ROI idiosyncrasies.

Conclusion

It appears that including the semantic vector, and creating a multimodal model increases performance slightly at inference time probably due to a subtle regularising effect, though the significance of the increase merits further study.

References

Allen, Emily J., Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, et al. 2022. "A Massive 7T fMRI Dataset to Bridge Cognitive Neuroscience and Artificial Intelligence." *Nature Neuroscience* 25 (1): 116–26. https://doi.org/10.1038/s41593-021-00962-x.

Gifford, A. T., B. Lahner, S. Saba-Sadiya, M. G. Vilas, A. Lascelles, A. Oliva, K. Kay, G. Roig, and R. M. Cichy. 2023. "The Algonauts Project 2023 Challenge: How the Human Brain Makes Sense of Natural Scenes." arXiv. https://arxiv.org/abs/2301.03198.

Haxby, J. V., M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. 2001. "Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex." *Science (New York, N.Y.)* 293 (5539): 2425–30. https://doi.org/10.1126/science.1063736.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. "Microsoft COCO: Common Objects in Context." arXiv. https://arxiv.org/abs/1405.0312.

McCarthy, Cormac. 2017. "The Kekulé Problem." *Nautilus*. https://nautil.us/the-kekul-problem-236574/.

Appendix