A Multimodel Approach to the Algonauts Challenge

Noah Syrkis & Sophia De Spiegeleire

June 6, 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, and training logs are available at wandb.ai/syrkis/neuroscope.

Introduction

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 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 (**lin2014?**) with fMRI responses to

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

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 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.

Literature review

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 ((Gonzales-Casillas2018?), (Groen, Silson, and Baker 2017)). 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), (Groen, Silson, and Baker 2017)). With the development of fMRI, it is now possible to explore and visualize real-time brain activity associated with visual perception ((Allen et al. 2022), (Haxby et al. 2001)). Despite this, the understanding of the intricate interconnectivity in visual perception, particularly for natural scenes, remains limited. Additionally, it is increasingly evident that to unravel the complex network underpinning visual perception, massive amounts of data are required ((Chang et al. 2019), (Allen et al. 2022)).

Multi modality in visual perception and deep learning

From a human-centered standpoint, multimodality pertains to the multiple sensory systems through which humans perceive and interact with the world ((Parcalabescu, Trost, and Frank 2021)). It is a reflection of the brain's capacity to integrate and process information from multiple sources. In the context of machine learning, multimodality refers to utilizing multiple information sources to enhance algorithms' performance. Several studies have pointed out the benefits of multimodal learning in providing richer information about underlying data patterns and creating more complex feature representations ((Ngiam et al. 2011), (Gu et al. 2017)).

Deep learning and visual encoding models

Deep learning has had a profound impact on neuroscience, particularly in understanding the brain's visual processing mechanisms. The success of deep learning models in neuroscience is attributed to their ability to process high volumes of data, their inherent flexibility, and their structure, which is inspired by the brain's own hierarchical organization ((kriegeskorte2015?), (kell2019?)).

In visual neuroscience, deep learning models have been extensively used to predict brain activity in response to visual stimuli, known as visual encoding models. A number of studies have demonstrated that deep neural networks (DNNs), particularly convolutional neural networks (CNNs), can accurately predict neural responses to various visual stimuli ((khaligh-razavi2014?), (yamins2014?), (cichy2016?)). Notably, these DNNbased models can even outperform traditional hand-engineered models ((guclu2015?), (kell2019?)).

However, most of these models rely on single-modal input data, typically the visual stimuli themselves. One study applied a convolutional recurrent neural network (CRNN) to investigate the computational mechanisms of the retinal circuit involved in interpreting natural scenes.The researchers discovered that recurrent spatiotemporal receptive fields of ganglion cells played a crucial role in encoding dynamic visual scenes.The findings also inciate that the inherent recurrence of the model enhanced the prediction the neural response, but also unveiled corresponding biological counterparts,emphasizing the power and potential of deep learning in visual neuroscience studies (Zheng et al. (2021)).

Zheng et al. (2021), applied a convolutional recurrent neural network (CRNN) to investigate the computational elements of the retinal circuit involved in interpreting the nature of natural scenes.Their findings highlight the instrumental role of the recurrent spatiotemporal receptive fields of ganglion cells in encoding dynamic visual scenes.The findings also inciate that the inherent recurrence of the model enhanced the prediction the neural response, but also unveiled corresponding biological counterparts,emphasizing the power and potential of deep learning in visual neuroscience studies.

Few studies have investigated the potential of multimodal deep learning for predicting fMRI responses. 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.

Han et al. (2019) aimed to investigate the use of

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, in accoradance with the Algonauts Project, 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 contains 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 (**krizhevsky2012?**). As in the Algonauts guide, PCA was performed reducing each image to a vector of size 100.

Over the course of 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.

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

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 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 focuses, 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.

Experiments

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 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 - α)((1 - β)$Lossy_{lh}$ + β $Lossy_{rh}$) + α $Loss_{y_c}$, 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 with respect to validation left hemisphere correlation in one sweep, and validation 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

In **table 1** we see the mean median voxel correlations for the two hemisphere versions of model 1 (the primary model with the auxiliary task) trained with and without α and β set to 0. The none baseline has $\alpha = 0.5$ and $\beta = 0.25$. **Table 1** shows us that our multimodal baseline outperforms the baseline on the test data. Further analysis would however be needed to explore the significance hereof. We also see that the *baseline* performs better on the train data. The baseline is thus, everything else being equal, overfitting more than its multimodal counterpart. This could be an indication that our semantic vector does indeed have a regularising effect.

With learned from the hyperparameter search described in **table 2** and **appendix A** and **appendix B**, we set α and β to 0.05 and 0.25 respectively.

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 the correlation between α and β the median correlation performance metrics. These correlations are based on a total of 60 runs across the five folds for each of the six subjects. The correlations are low, indicating that the usefulness of predicting the semantic vector as an auxiliary task, is at best subtle and at worst spurious. Again more elaborate statical analysis would be needed.

Table 2: Bayesian Hyperparameter Sweep (Model 1).

Hemisphere	α correlation	β correlation
Left	0.063	- 0.147
Right	0.076	- 0.087

In **table 3** we see the mean median voxel correlations across all subjects and folds of model 2 with (Alex + COCO) and without (Alex) 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 3: 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

A mean (across all subjects and folds) median voxel correlation projection onto a common cortical atlas is available interactively at neuroscope.streamlit.app/.

Analysis and Discussion

Future Work

As seen in the Analysis and Discussion, 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, 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, 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." January 10, 2023. http://arxiv.org/abs/2301.03198.

Han, Kuan, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. 2019. "Variational Autoencoder: An Unsupervised Model for Encoding and Decoding fMRI Activity in Visual Cortex." *NeuroImage* 198 (September): 125–36. https://doi.org/10.1016/j.neuroimage.2019.05.039.

Zheng, Yajing, Shanshan Jia, Zhaofei Yu, Jian K. Liu, and Tiejun Huang. 2021. "Unraveling Neural Coding of Dynamic Natural Visual Scenes via Convolutional Recurrent Neural Networks." *Patterns* 2 (10): 100350. https://doi.org/10.1016/j.patter.2021.100350.

Appendix

Appendix A

Table 4: lh sweep log (sorted by lh corr).

Name	alpha	beta	**val_lh_corr**	val_rh_corr
giddy-sweep-48	0.031	0.347	0.245	0.251
avid-sweep-12	0.051	0.223	0.245	0.23
woven-sweep-13	0.046	0.082	0.242	0.215
solar-sweep-11	0.076	0.167	0.24	0.226
leafy-sweep-18	0.025	0.36	0.235	0.253
dainty-sweep-44	0.044	0.018	0.229	0.224
effortless-sweep-47	0.033	0.042	0.228	0.209
fresh-sweep-17	0.034	0.072	0.225	0.177
silvery-sweep-41	0.096	0.218	0.225	0.225
fresh-sweep-50	0.014	0.211	0.224	0.209
generous-sweep-5	0.072	0.294	0.22	0.22
autumn-sweep-16	0.026	0.469	0.218	0.218
crimson-sweep-15	0.08	0.41	0.215	0.202
skilled-sweep-45	0.049	0.186	0.215	0.201
sandy-sweep-4	0.057	0.473	0.214	0.177
fearless-sweep-7	0.031	0.398	0.213	0.212
fanciful-sweep-20	0.042	0.429	0.212	0.22
sunny-sweep-14	0.029	0.353	0.211	0.2
genial-sweep-8	0.07	0.488	0.211	0.195
colorful-sweep-43	0.017	0.158	0.206	0.186
hardy-sweep-57	0.007	0.126	0.202	0.198
glamorous-sweep-42	0.025	0.201	0.201	0.185
dark-sweep-2	0.012	0.391	0.196	0.18
likely-sweep-46	0.056	0.067	0.194	0.184
pleasant-sweep-51	0.051	0.004	0.193	0.171
pious-sweep-31	0.065	0.347	0.19	0.223
cool-sweep-19	0.027	0.152	0.19	0.223
hearty-sweep-1	0.073	0.045	0.189	0.166
resilient-sweep-60	0.049	0.061	0.184	0.175
super-sweep-10	0.015	0.437	0.183	0.176
bright-sweep-34	0.065	0.066	0.181	0.203
light-sweep-23	0.092	0.191	0.177	0.165
firm-sweep-28	0.044	0.117	0.177	0.172
volcanic-sweep-49	0.0	0.412	0.174	0.179
breezy-sweep-6	0.054	0.451	0.173	0.165
tough-sweep-22	0.027	0.058	0.173	0.149

Name	alpha	beta	**val_lh_corr**	val_rh_corr
icy-sweep-9	0.085	0.373	0.173	0.157
distinctive-sweep-30	0.083	0.084	0.168	0.166
decent-sweep-58	0.09	0.499	0.159	0.148
true-sweep-3	0.042	0.354	0.157	0.144
firm-sweep-59	0.025	0.31	0.155	0.155
sweepy-sweep-56	0.075	0.27	0.153	0.112
clean-sweep-27	0.021	0.179	0.151	0.126
fallen-sweep-35	0.042	0.078	0.147	0.163
celestial-sweep-36	0.088	0.244	0.146	0.146
golden-sweep-55	0.015	0.396	0.146	0.131
splendid-sweep-53	0.048	0.07	0.143	0.149
efficient-sweep-26	0.02	0.387	0.143	0.135
fiery-sweep-54	0.04	0.442	0.138	0.139
firm-sweep-37	0.056	0.275	0.136	0.131
rosy-sweep-38	0.049	0.395	0.134	0.148
wandering-sweep-32	0.025	0.249	0.132	0.149
twilight-sweep-40	0.035	0.089	0.127	0.132
glorious-sweep-33	0.021	0.422	0.126	0.135
dainty-sweep-52	0.008	0.283	0.124	0.108
peach-sweep-25	0.029	0.401	0.115	0.116
devoted-sweep-21	0.093	0.412	0.11	0.1
true-sweep-24	0.04	0.259	0.107	0.111
fancy-sweep-39	0.004	0.083	0.106	0.105
radiant-sweep-29	0.061	0.427	0.104	0.106

Appendix B

Table 5: rh sweep log (sorted by rh corr).

Name	alpha	beta	val_lh_corr	**val_rh_corr**
wobbly-sweep-19	0.059	0.234	0.273	0.28
generous-sweep-49	0.089	0.045	0.234	0.237
deep-sweep-9	0.067	0.083	0.216	0.228
treasured-sweep-16	0.088	0.379	0.221	0.228
quiet-sweep-11	0.014	0.18	0.225	0.224
radiant-sweep-50	0.003	0.429	0.236	0.223
splendid-sweep-7	0.07	0.464	0.236	0.223
fallen-sweep-48	0.057	0.486	0.243	0.222
genial-sweep-42	0.038	0.037	0.226	0.221
misty-sweep-14	0.053	0.224	0.227	0.22

Name	alpha	beta	val_lh_corr	**val_rh_corr**
rosy-sweep-17	0.084	0.155	0.215	0.219
sweepy-sweep-34	0.078	0.333	0.193	0.218
ancient-sweep-18	0.007	0.011	0.22	0.216
major-sweep-45	0.068	0.452	0.223	0.215
peach-sweep-12	0.072	0.292	0.222	0.209
winter-sweep-47	0.018	0.198	0.211	0.207
warm-sweep-25	0.052	0.326	0.203	0.206
chocolate-sweep-46	0.01	0.103	0.202	0.205
crisp-sweep-20	0.095	0.02	0.19	0.203
effortless-sweep-44	0.03	0.33	0.236	0.201
warm-sweep-41	0.084	0.058	0.212	0.198
deep-sweep-10	0.02	0.331	0.21	0.196
kind-sweep-4	0.028	0.317	0.207	0.19
hearty-sweep-60	0.057	0.383	0.192	0.186
solar-sweep-1	0.034	0.302	0.192	0.186
absurd-sweep-43	0.003	0.129	0.206	0.186
dry-sweep-58	0.064	0.475	0.169	0.186
drawn-sweep-6	0.028	0.346	0.184	0.185
decent-sweep-35	0.097	0.225	0.167	0.184
worldly-sweep-5	0.042	0.396	0.172	0.183
kind-sweep-54	0.08	0.405	0.175	0.182
olive-sweep-31	0.022	0.222	0.158	0.181
sparkling-sweep-3	0.033	0.306	0.194	0.18
cosmic-sweep-37	0.079	0.143	0.155	0.177
light-sweep-13	0.058	0.253	0.204	0.177
devoted-sweep-57	0.08	0.167	0.177	0.175
proud-sweep-8	0.025	0.436	0.197	0.174
rural-sweep-53	0.016	0.423	0.192	0.173
earnest-sweep-2	0.071	0.16	0.178	0.172
visionary-sweep-36	0.075	0.223	0.16	0.171
floral-sweep-30	0.075	0.433	0.169	0.165
super-sweep-38	0.048	0.035	0.155	0.165
astral-sweep-28	0.044	0.285	0.168	0.163
neat-sweep-33	0.026	0.006	0.16	0.162
jolly-sweep-40	0.067	0.213	0.15	0.162
firm-sweep-59	0.002	0.3	0.15	0.16
cerulean-sweep-15	0.055	0.489	0.162	0.159
lilac-sweep-27	0.007	0.472	0.177	0.157
glad-sweep-56	0.085	0.451	0.16	0.156
dandy-sweep-32	0.089	0.048	0.138	0.151
glad-sweep-52	0.06	0.463	0.156	0.149

Name	alpha	beta	val_lh_corr	**val_rh_corr**
valiant-sweep-22	0.04	0.096	0.13	0.143
cerulean-sweep-51	0.065	0.208	0.142	0.143
still-sweep-39	0.046	0.062	0.125	0.142
smart-sweep-23	0.023	0.036	0.132	0.139
youthful-sweep-55	0.009	0.249	0.134	0.138
fresh-sweep-29	0.02	0.416	0.108	0.129
driven-sweep-26	0.079	0.167	0.133	0.125
autumn-sweep-24	0.079	0.484	0.125	0.109
skilled-sweep-21	0.041	0.338	0.069	0.086