A Multimodel Approach to the Algonauts Challenge

Noah Syrkis & Sophia De Spiegeleire

June 5, 2023

some title ideas:

- Integrating Multimodality in Deep Learning Visual Encoding Models: A Study on Natural Scenes Dataset

Abstract

Developing a computational model of how the brain decodes visual information is an important goal in neuroscience. In this project, focus on improving the encoding model of the Algonauts Challenge. Our approach, rather than deepening the model, is to add a modality during training. Specifically, we add a vector of semantic features to image shown to the subject as the fMRI data is collected. We find that this improves the performance of the model.

Introduction

first section (general introduction)... [importance of problem, short study aim]

[**general into**; studying the tech] The intricate nature of the human brain, often likened to a complex network of interconnected neurons, allows us to unravel the mysteries of human cognition. While its complexity remains a challenge in understanding human cognition in its full...

[**visual info processing**] Understanding how the brain encodes visual information is an important goal in neuroscience,....... While there has been significant research done investigating how the human brain encodes and processes visual information, there is still a vast amount that remains to be fully understood

- [**general brain**; studying the brain tech] The intricate nature of the human brain, often likened to a complex network of interconnected neurons, allows us to unravel the mysteries of human cognition. While its complexity remains a challenge in understanding human cognition in its full...

- The brain, a complex organ, can be conceptualized as a network of neurons. its properties are also physical properties that can be measured, on a more coarse level—in the case of this project, using fMRI. —> but new methods are needed!! to better understand what is goin go. . . .
  - big data: 'understanding complex networkd will inevitably reuqire massive amounts of data'Allen et al. (2022)

[multi modality???]

[short overview of research]

(write ltr)

–> now lit review

- Previous work
- Summarize existing approaches/methodologies
- Highlight research gaps

Visual information processing

- aim of section: explain the significance of visual information processing and its role in neuroscience research, traditional hierarchical models of visual processing, highlight the complexity and interconnectivity of visual processing, especially in the context of scene perception (most realistic), emphasize the need for comprehensive datasets, like the NSD, to understand the human visual system

  –

- [**vision**] While significant strides have been made in understanding these mechanisms, there remains a vast amount yet to be fully comprehended.

- Visual information processing is the principal channel for gathering information from our environment and lies at the forefront of neuroscience research. As the most studied sensory system, it plays a vital role in shaping our understanding a part of the intricacy of human perception ((**gonzales-casillas2018?**)). The mechanisms through which the brain encodes and processes visual stimuli have been the focus of significant research, adding to . . . .. However, the complexity of the visual information processing system suggests there is still a vast amount that remains to be fully understood.

- how does [fMRI] help us study this?

  - Functional magnetic resonance imaging (fMRI) has been groundbreaking (search syn) in shedding light on the neural correlates of visual perception, enabling the acquisition, analysis, mapping, and visualization of real-time brain activity through measuring blood-oxygen-level-dependent (BOLD) responses throughout the whole brain. (**allen2023?**), Haxby et al. (2001)

- While the brain is a complex organ, perhaps most accurately conceptualized as a network of neurons, it is also a physical whose properties can be measured, on a more coarse level—in the case of this project, using fMRI.
- [for later = ability to extract huge data, whole brain, big data]

The 'multi-modality' of visual information processing What is shown? [visual neuroscience literature]

- [traditional view: shorter] Traditional models of visual processing have often been categorized into three hierarchical levels: low, mid, and high-level visual information processing. Low-level processing focuses on elementary visual features such as lines, edges, color and constrast, while mid-level processing involves conjunctions of elementary features resulting in perceptual phenomena such as geometric primates and surface texture features, for example. High-level processing involves the abstraction of visual input into categorical or semantic representations for classification and identification ((**gonzales-casillas2018?**), (**groen2017?**)).

- [traditional view: longer (+ brain structure)] Traditional models of visual processing have often been categorized into three hierarchical levels: low-, mid-, and high-level stages. Low-level processing, occurring in the striate cortex (V1), engages with basic visual features such as line orientation (e.g. classic line orientation study (**hubel&wiesel1968?**)), edges, color, and contrast while transmitting more complex visual information to higher brain areas. The mid-level stage involves the integration of these elementary features, with significant contributions from the inferotemporal cortex (ITC) in the temporal lobe. This area processes diverse visual information such as object discrimination and representation, and connects extensively with memory and lower-level visual areas, aiding these processes ((**gonzales-casillas2018?**)). The high-level stage involves the abstraction of visual input into categorical or semantic representations for classification and identification. It's at this stage where memory, both semantic and working, comes into play, retrieving and reinforcing information from the ITC and storing new visual features when needed ((**groen2017?**)). [structure?]

- However, the categorization of visual features into these levels, while a useful framework for understanding visual information processing, is not all-encompassing. Category boundaries can be blurred and may overlap depending on the context or the required task. This ambiguity becomes apparent when comparing various tasks, such as edge detection, which relies primarily on low-level visual processing, for example of line orientations (in v1; source) compared to object detection or recognition tasks. Interestingly, the latter requires overlap and the integration of features at multiple levels of processing ((**groen2017?**)). [more info about object recognition]

    - Haxby et al. (2001) "a distinct pattern of response was found for each stimulus category" but also "representations of faces and objects in the ventral temporal cortex are widely distributed and overlapping" (faces, cats, man-made objects..) BUT 'small/submaximal responses are key' —> 'masked/covered categories = category could still be identified from fMRI data (p.2427)
    - (**clarke2011?**) object recognit– MEG… "showed increased recurrent interactions as a function of semantic integration demands… cortical dynamics of object processing are modulated by the complexity of semantic information required from the visual input."
    - (**kliger&yovel2019?**)

- In the context of **scene visual perception**, compared to object recognition, the contributions of feature processing appear to be even more intertwined, challenging the hierarchical division suggested by traditional models of visual processing. In general, these models fall short in adequately representing the complexity of real-world scene perception, which is highly dynamic, involving integrating information from multiple objects and identifying spatial and functional relationships as well as scene gists and categories, for example ((**groen2017?**)). Thus, it is proposed that future research should foucs on comprehending the contribution of various scene properties as well as the interconnectivity of neural mechanisms and pathways regarding the visual information processing of natural scenes ((**groen2017?**)).

    - The visual processing of natural scene images, in addition to being ecologically relevant, are effective activators of the entire (visual) system" (Allen et al. (2022) – see citation12),
    - V1/v2 –> "recurrent interactions in visual cortex between areas along the ventral stream and striate cortex play a causal role in categorization and perception of natural scenes" (**koivisto2011?**)

- Current perspectives suggest that visual information processing involves a sophisticated interplay of interconnectivity and 'multimodality' (complex architecture). Allen et al. (2022) ; our ability to efficiently perceive the visual world is underpinned by a remarkably interconnected and multimodal network, comprising approximately one-third of the human cerebral cortex and interconnects brain regions with various different functional properties (add to this - citation 7 and 8). Allen et al. (2022); This network functions not only to encode visual stimuli but also integrates visual representations into a cognitive context, influenced by what one has previously seen, might see, or is selectively attending. (add – citation 9, 10, 11) .. (Allen et al. (2022) – see citation12),

- Problem 1**: we don't know that much about this interconnectivity/complexity for Natural scenes** = complex stimuli = complex visual information processing —– and to understand this we need LOTS of data

- that is why we need to study "whole-brain responses to complex stimuli critical in the quest to understand the human visual system." Allen et al. (2022) … (fmri)
- big data: 'understanding complex networks will inevitably require massive amounts of data'Allen et al. (2022) —> especially in visual neuroscience ""Neuroscience has an insatiable appetite for data""
- great paper for this problem (**chang2019?**) "how do we scale neural datasets to be commensurate with state-of-the-art computer vision, encompassing the wide range of visual inputs"
  - "neural datasets studying biological vision are typically lacking in: (1) size; (2) diversity; and (3) stimulus overlap; relative to extant computer vision datasets" (**chang2019?**)
- The NSD data set – "the dataset will be useful for investigating a variety of phenomena in low-, mid-, and high-level vision." and the interconnectivity between them / study them as a whole (but this is difficult for humans)
- interdisciplinary approaches keep developing [transition to DL vision encoding models]

Deep learning + visual encoding models

- aim of section: explain the impact of the deep learning revolution on neuroscience research (visual encoding models), explain how deep learning algorithms have been applied to fMRI data analysis, emphasize the need for large-scale fMRI datasets, multimodality in deep learning models to tackle the complexity of the brain
- [+ deep learning revolution + neurosci= visual encoding models]
  - Over the course of the last decade, the deep learning revolution has had a profound impact on scientific research endeavors in neuroscience!!!!! With the ability to process substantial volumes of data and uncover intricate patterns, deep learning algorithms have revolutionized our understanding of the brain.
  - (**chang2019?**) "high-performing computer vision systems have been touted as effective potential models of neural computation – This is primarily for three reasons: (1) the origin of these models is linked to the architecture of the primate visual system4; (2) these models learn from millions of real-world images; (3) these models achieve high-performance in diverse tasks such as scene recognition, object recognition, segmentation, detection, and action recognition tasks"
    - CNN – "use network layers weights to predict neural responses at the voxel (fMRI) or neuron (neurophysiology) level within that neural region… higher-level layers = predict higher-level object and scene regions" (possibly good voor discussion) (**chang2019?**)
  - (**mehrer2021?**) ecoset dataset ('ecologically relevant images'), NN AlexNet

= "training on ecoset leads to significant improvements in predicting representations in human higher-level visual cortex. . . .". . . . computational visual neuroscience should use image sets that reflect the human perceptual and cognitive experience – natural scenes especially complex
  - These algorithms and models are inspired by the complex architecture of the brain itself . . . .. Gifford et al. (2023, neural networks — neuroimaging) have tried to tackle been applied to tasks such as fMRI data analysis, brain connectivity mapping, and image reconstruction and **
  - (**gu2017?**) - propose a multi-fusion deep learning framework that learns multimodal features richer in semantic - multi-modal fusion (text) *** good paper for multi-modal relevance and architectures
- visual **encoding** models; "voxel-based = predict activity in single voxels that are evoked by different sensory.. tasks = provide an explicit quantitative description of how information is represented in individual voxel activity" (**naselaris2011?**) Han et al. (2019)
  - components of encoding models: stimuli, features, ROIs (in separate spaces), input space, feature space, activity space– see Kay et al. ((**naselaris2011?**) p.401)
- experiment specific: predicting human visual brain responses through computational models
  - **** where do we add the 'experiment (encoding/decoding task) specific literature??' (**naselaris2010?**) (encoding and decoding firm)
  - "Encoding models have one great advantage over decoding models. A perfect encoding model provides a complete functional description of a specific ROI but a perfect decoding model does not." (**naselaris2010?**)
  - **Decoding** images from brain activity is a well-studied problem in the field of neuroscience. The first successful decoding of images from brain activity was done by Haxby et al. (2001). Like the current project, Haxby et al. used fMRI data. Most recently Lin, Sprague, and Singh (2022) used a deep neural network to decode images from brain activity. Also Thomas, Ré, and Poldrack (2023) merits mention, focusing on developing a mapping between brain activity and mental states more broadly.
- **Problem 2**: we need BIG DATASETS — fMRI mostly small (solution = NSD)
  - The scarcity of large-scale fMRI datasets hinders the full potential of deep learning approaches, emphasizing the need for comprehensive datasets like the Natural Scenes Dataset (NSD)!
- **Problem 3**: if the human brain learns multi modally, shouldn't deep learning algorithms as well? how can we do this? we need STUDY visual perception multi modally — (accurately model complex brain structures)
  - utilize Richer (meta)data —— e.g. language, 'semantic features' (categories)

Natural Scenes Dataset + Algonauts

- maybe incorporate this into next sub-section 'Research questions and aim/hypotheses'
- NSD Created to solve all these 'research gaps' (problems (1,2... kind of 3 also)) (!!!!)
  - We have the data ... now we need to create models + test it + add multimodality
  - considering we have this much data now –> better for DL vision encoding/decoding models
- We use a subset of the Natural Scenes Dataset (NSD) Allen et al. (2022), provided by the Algonauts Project Gifford et al. (2023), to train a model that, given an image, can predict the fMRI response of a subject.

  The Algonauts Project is a competition that aims to develop a computational model of how the brain encodes visual information. This is foundational research with potential applications in both neuroscience and machine learning.

  The NSD consists of 73,000 images of natural scenes and various assoicated responses, collected over the course of one year from 8 subjects, making it the largest dataset of its kind, enabling the development of more accurate models, which are now released on an ongoing basis from various research groups.

Research questions and aim/hypotheses

- In light of these considerations (research gaps) our research aims to integrate an additional modality, through a vector of semantic image features, for a deep learning encoding model...
- using the LARGEST fMRI natural scenes dataset currently available –>
- We aim to explore the research questions: How can we effectively incorporate multimodality (computer vision and semantic features (symbolic)) in our deep learning visual information encoding model? What are the implications of including semantic features during training in the encoding model for a decoding task (image to predicting brain activity (fMRI) and/or encoding task (fMRI to visual features)?
- Our approach is to add a modality during training. Specifically, we add a vector of semantic features of the image for the given fMRI data. Thouch multimodality, is a common approach in machine learning, recent advances in deep learning has largely been enabled by enormous amounts of data.
- Research question and hypotheses
- Structure of paper

Our approach is to add a modality during training. Specifically, we add a vector of semantic features of the image for the given fMRI data. Thouch multimodality, is a common approach in machine learning, recent advances in deep learning has largely been enabled by enormous amounts of data. In this project, we explore the potential of multimodality in the context of the Algonauts Challenge, attempting a model that, during inference, has the same number of parameters as the unimodal baseline model.

Methodology

- Introduction
- Data collection (describe datasets in more detail)
- Feature selection, model architecture
- Describe experiment(s)
- Describe analysis and evaluation metrics

Dataset

The dataset is provided by the Algonauts Project Gifford et al. (2023). It is originally sourced from the Natural Scenes Dataset (NSD) (Allen et al. (2022)), currently the largest dataset containing cortical surface vertices from the brain in both the left and right hemispheres of eight participants in response to various images depicting natural scenes. The study utilizes 73,000 images of natural scenes sourced from the Microsoft Common Objects in Context (COCO) dataset ((**lin2014?**)). Each participant in the NSD study viewed 10,000 unique images over the course of one year. Every image was presented three times, culminating in 30,000 image trials per subject. The images are 254x254 pixels, and the fMRI data for the left and right hemispheres are 19,004 and 20,544 voxels respectively. These vertices were determined to be maximally responsive to the images based on preprocessed, high-quality 7T fMRI responses, expressed as BOLD response amplitudes. Furthermore, associated with each subject are region of interest (ROI) masks. These masks are used to extract the fMRI data from the images, at specific locations in the brain.

In our research, we focus exclusively on data from the left hemisphere of six selected subjects, chosen for their consistent voxel counts in both hemispheres, ensuring a balanced comparison across subjects. Apart from this specific selection, the NSD fMRI data is utilized in its original preprocessed form as provided by the Algonauts Project, maintaining its preprocessed state to ensure consistency and reliability in its application.

Model architecture + (multi-modal) feature selection

In accordance with the Algonauts Challenge, compute each image's AlexNet features (**krizhevsky2012?**), and uses these as the input to the model. Specifically, we use the features from the fc2 layer of the network, so as to keep our baseline as close to the Algonauts Challenge baseline as possible. In accordance with that baseline, we also perform principal component analysis (PCA), reducing the dimensionality of the features to 100.

Vectors of semantic features are provided by the Common Objects in Context (COCO) dataset (**lin2014?**). The COCO dataset consists of 328,000 images of 91 object categories, 80 of which are present in the NSD. Thus a sample of our data consists of the four-tuple (image, left fMRI, right fMRI, semantic features). The semantic features are 80-dimensional vectors, one for each object category. An image can contain multiple objects, so most semantic vectors contain multiple ones.

- components of encoding models: stimuli, features, ROIs (in separate spaces), input space, feature space, activity space– see Kay et al. ((**naselaris2011?**) p.401)
- (**mocz2023?**) 'Representing Multiple Visual Objects in the Human Brain and Convolutional Neural Networks' = check dis (abt 'individual voxel averaging')

Experiments

Our experiment, attempting to improve performance through multimodality without increasing inference parameters, tests the effect of predicting both the fMRI response and the semantic features from the image during training. The input is thus always the image (represented as the AlexNet features), and the output is either the fMRI response (during pure inference), or both the fMRI response and the semantic features (during training). Hopefully, this will allow the model to learn a more accurate representation of the image, and thus improve performance, on the fMRI response, during inference, without increasing the number of parameters.

We use K-fold cross-validation, with K=5, to evaluate the performance of our model, using the Pearson correlation coefficient as the metric. Our loss function is the mean squared error (MSE) between the predicted and actual fMRI response. We use the same model architecture for every subject and hemisphere, though the parameters are reinitialized for each subject-hemisphere pair.

The model architecture is a simple feedforward neural network, as our focus is on multimodality, rather than deepening the model or exploring other architectures. The model has two outputs, one for the fMRI response, and one for the semantic features. From a neuroscience perspective, the fact of us predicting the blood oxygenation level-dependent (BOLD) signal, for multiple regions of interest (ROIs), is already multimodal, as the ROIs are in different parts of the brain, and function by vastly different mechanisms.

Results

-

Discussion

- Explanation of interpretation of main results
- Compare/link to the literature
- Discussion of strengths and limitations of the approach
- Proposal of perspectives for future research

limitations

- fMRI limitations see (**qian2023?**) (preprint);"" fMRI signals are an indirect measure of brain activity, the information they contain may not always correlate with the semantics of the observed image"

further research

- Semantic Neural Decoding via Cross-Modal Generation (**qian2023?**) (preprint) – (decoding) SemanSig (deep generative model to decode the semantic information), creates class prototype space as the internal representation space of fMRI signals
- further develop encoder-decoder model; "A model that truly reflects the brain's algorithmic mechanism of vision should be image-computable and capable of predicting brain responses to any visual input (namely encoding) and retrieving visual and conceptual information from brain responses (namely decoding)." Han et al. (2019) ; p.407 ; (**Naselaris2011Georgopoulos?**) et al. (1986)
- (**gu2017?**) - multimodal – add text/caption; paper propose a multi-fusion deep learning framework that learns multimodal features richer in semantic - multi-modal fusion

Conclusion

- Short overview – summary of main results
- Reinforce the significance of results and (potential) impact

References

- APA 7th, (in-text citations)

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.

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.

Lin, Sikun, Thomas Sprague, and Ambuj K. Singh. 2022. "Mind Reader: Reconstructing Complex Images from Brain Activities." September 30, 2022. http://arxiv.org/abs/2210.01769.

Thomas, Armin W., Christopher Ré, and Russell A. Poldrack. 2023. "Benchmarking Explanation Methods for Mental State Decoding with Deep Learning Models." *NeuroImage* 273 (June): 120109. https://doi.org/10.1016/j.neuroimage.2023.120109.

Appendix