Predicting fMRI Volumes from Natural Scene Images and Text Descriptions

Tom Fizycki
Ecole Polytechnique
tom.fizycki@polytechnique.edu
January 11th 2025

Abstract

The development of Computer Vision methods, be it purely deterministic or rather based on Deep Learning approaches, enabled for further understanding in the machine processing of images. With interpretability advances such as [7, 9], we are getting better at getting how exactly state-of-the-art black-box algorithms produce such impressive results. Nonetheless, looking at neuroscience's latest discoveries, we still lack understanding in how the information is transmitted from an image to the actual brain activity. To this end, [6] introduces an artificial neural network framework, VISION, for "Visual Interface System for Imaging Output of Neural activity", whose goal is to mimic human brain activity corresponding to the action of vision. Using as an input both natural scene images and a contextual description, this model attempts to predict brain's functional magnetic resonance imaging (fMRI) scan response.

This project aims at reproducing [6]'s promising results on the Natural Scenes Dataset (1), as well as proposing extensions for both performance improvement and insights. Though performances were nor matched nor improved, this report aims at describing one's journey through research on understanding how the human brain processes images.

1. Introduction

1.1. Background on fMRI

As explained in [4], the complex nature of visual stimuli makes the encoding of brain-relevant features much more complicated than it is for 1D signals. In order to focus on the perception of the stimuli, researchers most often focus on fMRI as the brain's approach to interpresting the contained information. As developpped in [8], when specific brain regions are activated in response to stimuli, there is an increase in oxygenated blood, which fMRI captures using the blood-oxygen-level-dependent (BOLD) signal. This makes it particularly valuable for studying the functional organization of the brain, such as the visual cortex, where it helps researchers map the response to visual inputs like shapes, colors, and motion.

1.2. The Natural Scenes Dataset (NSD)

The Natural Scenes Dataset (NSD) [1] comprises high-resolution functional magnetic resonance imaging (fMRI) data acquired using a 7T scanner, focusing on the brain's responses to tens of thousands of richly annotated natural scene images.

Data Available: NSD includes fMRI data collected from eight participants across 30–40 MRI scanning sessions, with each session lasting approximately 7.5 hours. The dataset provides over 22,000 unique scene presentations per participant, encompassing repeated stimuli to enhance neural signal-to-noise ratio. Alongside fMRI data, the dataset offers extensive behavioral annotations, resting-state fMRI, diffusion MRI, and structural scans to facilitate network-level analysis of brain activity data manual.

Data Collection: The NSD was acquired while participants performed a continuous recognition memory task, observing images from the COCO dataset—a widely used corpus of labeled natural images. Participants provided behavioral responses during these trials, enabling simultaneous neural and cognitive assessments. To ensure data quality, state-of-the-art denoising and estimation techniques were applied, yielding high signal fidelity, particularly along the ventral visual pathway [1].

2. Methodology

The method used here relies on the architecture presented in [6].

2.1. Multimodal Neural Encoding Model

The proposed framework, VISION, integrates two principal components: a multimodal feature extractor and a dense-channel encoding interface network. These modules aim to replicate the visual cortex's processing of both semantic and visual features, as shown in Figure 1.

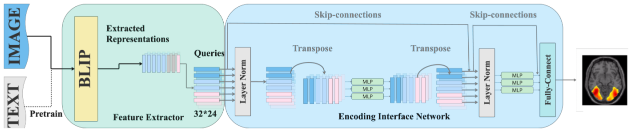


Figure 1. Base model architecture presented in [6]

Feature Extractor: The feature extractor is inspired by the dual function of the visual cortex in processing semantic and visual contextual information. The state-of-the-art pretraining model, BLIP [5], is employed for this purpose. BLIP incorporates a vision transformer model [3] and three transformer-based modules with architectural similarities to BERT [2]. Unlike its standard configuration, only image inputs are provided to BLIP to minimize computational overhead. The resulting BLIP features capture high-dimensional semantic and textual information derived during pretraining. These features are subsequently transformed into a structured query matrix, preserving their complex, layered organization.

Encoding Interface Network: To mimic the structure of the brain's neuronal networks, the encoding interface is based on a Multilayer Perceptron (MLP). Each MLP module comprises two fully connected layers. BLIP features are reshaped into a $32 \times 32 \times 24$ matrix, producing 768 independent queries, each of which is processed by the MLP. The outputs are then aggregated to represent the dense neural encoding of the feature space, as illustrated in Figure Figure 1.

2.2. Accuracy Metric: Noise-Normalized Accuracy

To evaluate model performance, the noise-normalized accuracy metric is employed, leveraging the noise ceiling value. Noise ceiling, calculated as shown in Equation 1, represents the maximum variance in voxel responses attributable to the signal under the presence of noise:

$$\text{NC} = 100 \times \frac{\sigma^2_{\text{signal}}}{\sigma^2_{\text{signal}} + \sigma^2_{\text{noise}}}. \tag{1}$$

Prediction accuracy is further defined using the correlation coefficient, R_v, between the ground truth (G_v) and predicted (P_v) fMRI responses of a voxel v, as shown in Equation 2:

$$R_v = \text{corr}(G_v, P_v). \tag{2}$$

The final accuracy is computed as the median noise-normalized correlation across all voxels, as shown in Equation 3.

$$\text{Accuracy} = \text{Median}\left(\frac{R_1^2}{NC_1}, \dots, \frac{R_v^2}{NC_v}\right) \times 100. \tag{3}$$

This metric ensures robust assessment of the model's predictive accuracy, incorporating noise considerations intrinsic to the dataset.

3. Implementation and experiments

In this project, we set up 3 different experiments in order to better grasp the proposed method, as well as the challenges that are raised by this exercise.

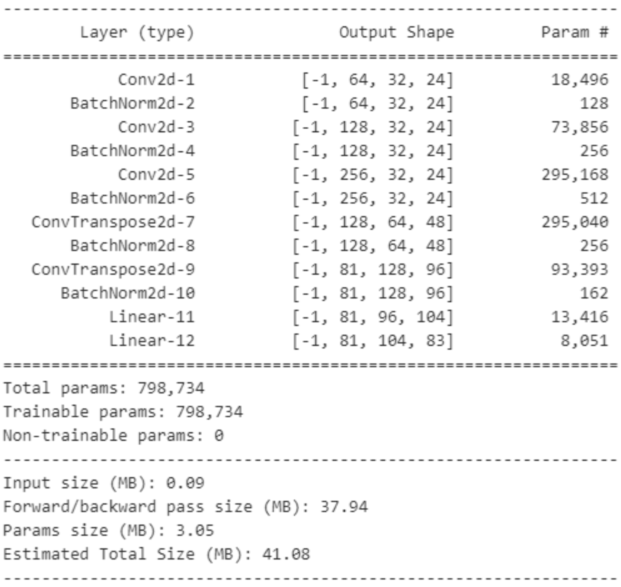
3.1. Reproduction of simple training

As a first experiment, we try to reproduce the results showcased in [6]. To this end, we thank the authors for the provided code in https://github.com/Rxliang/VISION, though its access was complicated considering the exact data used was not specified. This is even more complicating considering the considerable number of variants and preprocessings of the data that are made available by the NSD.

Amongst the used Encoding Interface Networks used for this experiment, two were used : the MLP-based one presented in Figure 1, as well as another architecture resorting periodic activation functions, such as presented in [10]. The latter claims this addition enables to extract complex features using implicit neural representations, with its associated code provided at https://github.com/vsitzmann/siren.

Due to the large size of the dataset, we were unable to perform training on GCloud VMs. Thereforte, the latter, as well as all expriments described below, were performed on CPU (12-core AMD Ryzen 5600H).

3.2. CNN-based Encoding Interface Network



```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1           [-1, 64, 32, 24]          18,496
       BatchNorm2d-2           [-1, 64, 32, 24]             128
            Conv2d-3          [-1, 128, 32, 24]          73,856
       BatchNorm2d-4          [-1, 128, 32, 24]             256
            Conv2d-5          [-1, 256, 32, 24]         295,168
       BatchNorm2d-6          [-1, 256, 32, 24]             512
   ConvTranspose2d-7          [-1, 128, 64, 48]         295,040
       BatchNorm2d-8          [-1, 128, 64, 48]             256
   ConvTranspose2d-9          [-1, 81, 128, 96]          93,393
      BatchNorm2d-10          [-1, 81, 128, 96]             162
           Linear-11          [-1, 81, 96, 104]          13,416
           Linear-12          [-1, 81, 104, 83]           8,051
================================================================
Total params: 798,734
Trainable params: 798,734
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.09
Forward/backward pass size (MB): 37.94
Params size (MB): 3.05
Estimated Total Size (MB): 41.08
----------------------------------------------------------------
```

Figure 2. CNN-based Encoding Interface Network

After this first experiment, whose results will be presented in section 4, it was decided to tryout a novel Encoding Interface Network. Its architecture is further developed in Figure 2. This experiment was inspired after observing non-desired simili-periodic predictions for the two Encoding Interface Networks presented in Experiment 1. This is normal considered the activation functions in [10].

3.3. Studying effects of penalization

In [6], authors highlight the strong sparsity of attention matrixes. This comes together with the very high sparsity observed by [11] inside the distribution of the human gaze regarding teh regions of an image to which the eye and the brain pay attention. To this end, we want to compare the results of adding penalization during training. We thus compare the results of L_1 and L_2 losses, the first being known to strongly incentivize sparsity in model weights, while the latter is known to push for smaller non-zero weights.

3.4. Experiment setup in details

In our experimental setup, the BLIP-based feature extractor is frozen, while the Encoding Interface Network is trained, in order to both ease training, and keep the already meaningful representations.

All the results were achieved with a $24000/2000/4000$ split for respectively training, evaluation, and testing datasets. Descriptions are padded/tuncated to 32 tokens. fMRI volumes are of size $81 \times 104 \times 83$ voxels. We used a 1pt 8mm resolution, as well as the fithrf, GLM-denoised version of the fMRIs.

Only one subject out of 8 was selected for practical purposes.

4. Results

4.1. Qualitative results

Considering the generative nature of our work, we shall first present a handful of qualitative results.

As for visualizing the results produced by our method and experiments, we showcase :

- An input example, including the effect of the built-in visual processor in Figure 3.
- The sucessive predictions after n training epochs of our method, compared to ground truth in Figure 4.
- A 3D view of both ground truth and one prediction, highlighting a different view in Figure 7.

4.2. Quantitative results

In Table 1, we display the numerical results obtained in our experiments, with the protocol described in section 3. As a reference, [6] obtains a score of 0.55 over all 8 subjects, averaging over brain vertices. The model of [4], evaluated in the previous paper achieves an average result of 0.17.

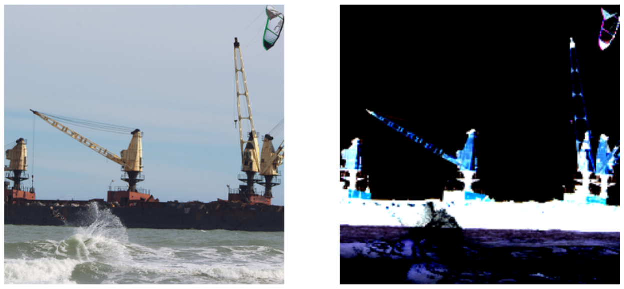


Figure 3. Example input, with the following description : *a group of boats with large giant cranes on top of them.* The figure shows the image before (on the left) and after (on the right) preprocessing. Here, we use the visual processor of the BLIP Feature Extractor used in the already provided code.

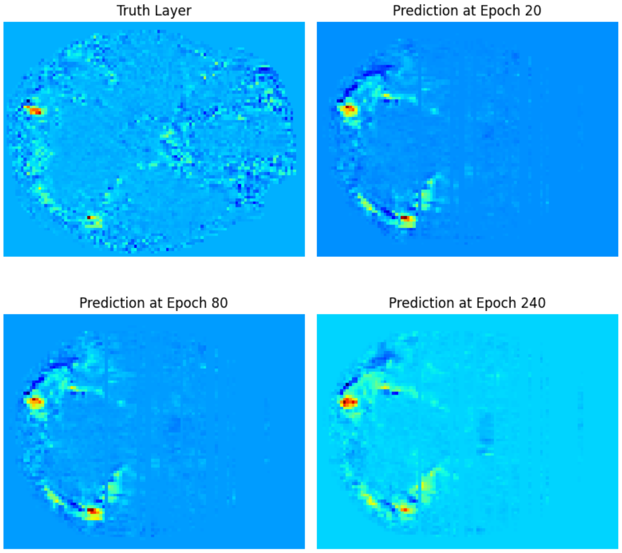


Figure 4. Example of output produced by our method on Figure 3. The chosen Encoding Interface Network is described in Experiment 2. We show only a certain horizontal cut of the brain (layer 30).

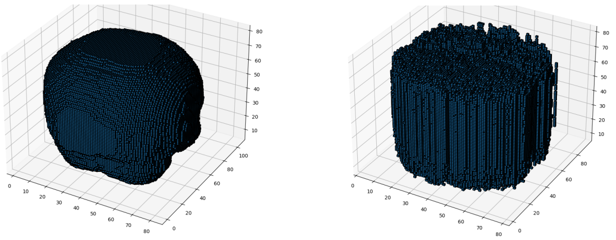


Figure 5. Truth 3D view.

Figure 6. Prediction 3D view.

Figure 7. Comparative 3D view of ground truth VS prediction on the same sample.

Regularization	Encoding Interface Network		
	MLP	**SIREN**	**CNN**
L1	0.12	0.16	0.07
L2	0.12	0.15	0.06

Table 1. Noise-Normalized accuracy for experiments described in section 3.

5. Discussion

5.1. Method capabilities and comparison with baseline results

When gazing at the results we try to replicate, we realise that we have far from achieved this goal. With an accuracy of 0.16, we fall shy to the 0.55 claimed by [6]. We might be reconciled when seeing we achieve results in the order of magnitude of [4].

On the other hand, when looking at the qualitative results, we are able to grasp a much better insight at our method's capabilities and failures. First, Figure 4 shows that, when looking at the middle horizontal layers of the brain, our model is able to produce distinct vertices. Moreover, it is accurate in noticing some high neural activity in vertices that are directly linked to the visualization of an image. With training, it also predicts larger activation zones, rather than simple pinpoints, which is much more realistic. In 3D, it comprehends what a brain roughly looks like, at the exception of the layers situated on top of the brain, as well as in the lowest, with a smallest error.

Taking a step back, it is quite undertrained, and is far from reaching the figures we hoped to replicate. A possible (though it remains an hypothesis) explanation of this is developed below.

5.2. Computational struggles

Going through this challenge, we leaped across a certain number of obsatcles. Firstly, the dataset was too large to write onto a GPU-accelerated GCloud VM, which compelled training to be performed on CPU, not making for a good start.

Furthermore, we had a lot of difficulty trying to replicate the experiment described in [6]. Though they indeed provide their code, they remain opaque regarding the necessary environment and requirements. Furthermore, the exact data used is not thoroughly explained by the authors, though the dataset itself is very well documented. To this end, we thank the authors of [1] for their immense contribution to the community.

It is possible, though not assured, that more time and a better computational organisation on our end, notably more capable GPUs for training might have improved significantly our performances.

6. Conclusion

Concluding this research project, we are happy to present our results, though unsatisfactory regarding the objectives we had at the start. Our experiments are backed by thoughtful thinking, and we stay persuaded they could very well lead to more insights towards understanding how our brain processes the task of vision. Taking a look back at our work, this was a very enriching experience, in multiple points. Diving into the feature extraction of VLMs, the neural processing of information, as well as the challenges raised by medical imaging, were all enlightening steps. We remain happy of our contribution, though much further analysis would be needed, and will gladly look back at everything we learned.

References

[1] A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. 2022. 1, 4

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019. 2

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*, 2021. 2

[4] Meenakshi Khosla, Keith Jamison, Amy Kuceyeski, and Mert Sabuncu. Characterizing the ventral visual stream with response-optimized neural encoding models. In *Advances in Neural Information Processing Systems*, pages 9389–9402. Curran Associates, Inc., 2022. 1, 3, 4

[5] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare, et al. Bootstrapped language-image pre-training (blip). *arXiv preprint arXiv:2201.12086*, 2022. 2

[6] Ruixing Liang, Xiangyu Zhang, Qiong Li, Lai Wei, Hexin Liu, Avisha Kumar, Kelley M. Kempski Leadingham, Joshua Punnoose, Leibny Paola Garcia, and Amir Manbachi. Unidirectional brain-computer interface: Artificial neural network encoding natural images to fmri response in the visual cortex, 2023. 1, 2, 3, 4

[7] Christopher Olah, Ludwig Schubert, and Alexander Mordvintsev. Feature visualization. *Distill*, 2017. 1

[8] Trinath T Logothetis NK Rainer G, Augath M. Nonmonotonic noise tuning of bold fmri signal to natural images in the visual cortex of the anesthetized monkey., 2001. 1

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”: Explaining the predictions of any classifier, 2016. 1

[10] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions, 2020. 2, 3

[11] Zijun Wei, Hossein Adeli, Minh Hoai Nguyen, Greg Zelinsky, and Dimitris Samaras. Learned region sparsity and diversity also predicts visual attention. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 3