Connecting pre-trained models with brain responses: Report for the Algonauts Project Challenge 2023

Vyacheslav Shen, Kassymzhomart Kunanbayev, Jeongwon Lee, and Daeshik Kim

Brain Reverse Engineering and Imaging Lab, KAIST

Abstract. This report encapsulates our participation in the Algonauts challenge 2023, offering an overview of the approaches and their outcomes. The challenge centered on the prediction of brain responses from the view of complex natural scenes. Our method was based on the application of different pre-trained large vision-language models to extract more comprehensive features of the images, which were linearly mapped into the fMRI responses. Our best submission scored 50.6 in the Noise-Normalized Mean Correlation (NNMC) metric. The code can be found at Github: `https://github.com/slavaheroes/breil-algonauts-challenge`.

1 Introduction

The primary objective of the Algonauts Project 2023 Challenge [1] revolves around the utilization of computational models to anticipate brain reactions that are recorded when participants observe natural visual scenes. The challenge was conducted in collaboration with the Natural Scenes Dataset (NSD) [2] which contains an extensive collection of brain responses associated with the scenes of the COCO dataset [3]. The NSD comprises of brain responses derived from 8 human participants, encompassing a total of 73,000 distinct visual scenes. These brain responses were measured through functional MRI procedures. The fundamental challenge objective is to forecast brain responses across the entire visual brain region.

2 Methods

To make Image-to-fMRI translation more efficient, we employed a range of large pre-trained deep-learning models to generate features for the images. Subsequently, our focus was on establishing a mapping between these embeddings and the corresponding fMRI responses. Through empirical experimentation, we discovered that the most effective approach for this mapping was achieved using linear regression with L2 regularization. The best hyperparameters for the linear model were found using grid search and cross-validation by comparing the Pearson correlation score.

Overall, the linear models were constructed for each hemisphere of each subject except for the submissions where we modeled a separate linear model for each region of interest (ROI). In the following subsections, we will mention the details of successful and unsuccessful methods.

2.1 Successful Approach

In current works, reconstructing images from fMRI responses [4,5,6,7] were accomplished using the pre-trained CLIP (Contrastive Language Pretraining) model [8] where fMRI responses were transferred to the CLIP embedding space. Motivated by these methods, we have also utilized CLIP models to generate the embeddings of the images and their corresponding COCO captions. Through submissions, we have concluded that CLIP-B [ViT-32] model was better than the CLIP-L model, so in the following experiments, we used the features generated by CLIP-B.

In [6], CLIP text embeddings were considered to have better semantic representations whereas CLIP image embeddings have good visual representations which can be mapped to different brain regions. Following the same logic, we experimented with the available large-vision pre-trained models in order to find a better visual representation. Among popular pre-trained vision models such as EVA [9], ViT [10], and Segment Anything (SAM) [11], the latter achieved a better score. In the case of SAM, only the image encoder part of the model, i.e. ViT-L with window attention, was used.

We conducted the analysis based on the per-ROI-based correlation score which became available after each submission. As shown in Fig 1, Segment Anything features exhibit a higher correlation in visual regions while CLIP image-only and concatenated image and text features demonstrate better correlation in other regions of the brain. We believe that it indicates that SAM has good visual representations while CLIP has better semantic representations of the given images because of their different training objectives. CLIP is tailored to learn joint representations of both images and text, which enables them to understand and capture the underlying semantics of visual content and associated language. This joint training encourages the model to extract features that can effectively express the relationships and meanings inherent in both visual and textual data. On the other hand, SAM is used for segmentation and optimized to extract features that are discriminative for distinguishing different object classes within images.

Best Submission Based on the analysis in Fig 1, linear models were constructed for each ROI to maximize the correlation within the corresponding regions. For example, for prf-visual ROI classes a linear model using only SAM features, for other regions a linear model using only concatenated CLIP image and CLIP text features was formulated. The schematics are shown in Fig 2. This approach improved the challenge score considerably as shown in Table 1, but empirically we found that a simple average of the SAM prediction and the

Fig. 1. NNMC score vs. ROI-classes

CLIP-B prediction for all ROIs resulted in a higher score which eventually was our best submission.

Other submission scores are shown in Table 1.

Table 1. Results of the noticeable submissions. **Bold** indicates the best score.

Features and Methods	NNMC score
CLIP-B image only	42.006
CLIP-B text only	36.591
CLIP-B image and text features concatenated	41.930
SAM Image Encoder	39.814
CLIP-B image and text concatenated + SAM features per ROI	48.006
CLIP-B image and text features + SAM features Neural Network	48.053
CLIP-B image and text + SAM features averaged	**50.632**

2.2 Unsuccessful Approaches

In this subsection, we would like to focus on the approaches that did not yield desired performance.

Diffusion Models Being inspired by the application of Diffusion Models in fMRI-Image translation [6,7,5], one of our approaches was to extract diffusion latent features. Unfortunately, the scores of this approach were very low.

MLP projector Neural Networks were considered as an alternative mapping method of features to the fMRI responses as in [4,5] where Neural Networks map fMRI responses to CLIP embedding space. We constructed a mapping network from SAM and CLIP embedding to the fMRI responses similar to [5], and trained it with Mean Squared Error and Cosine Distance losses. However, as shown in Table 1, this approach did not perform better than our best submission.

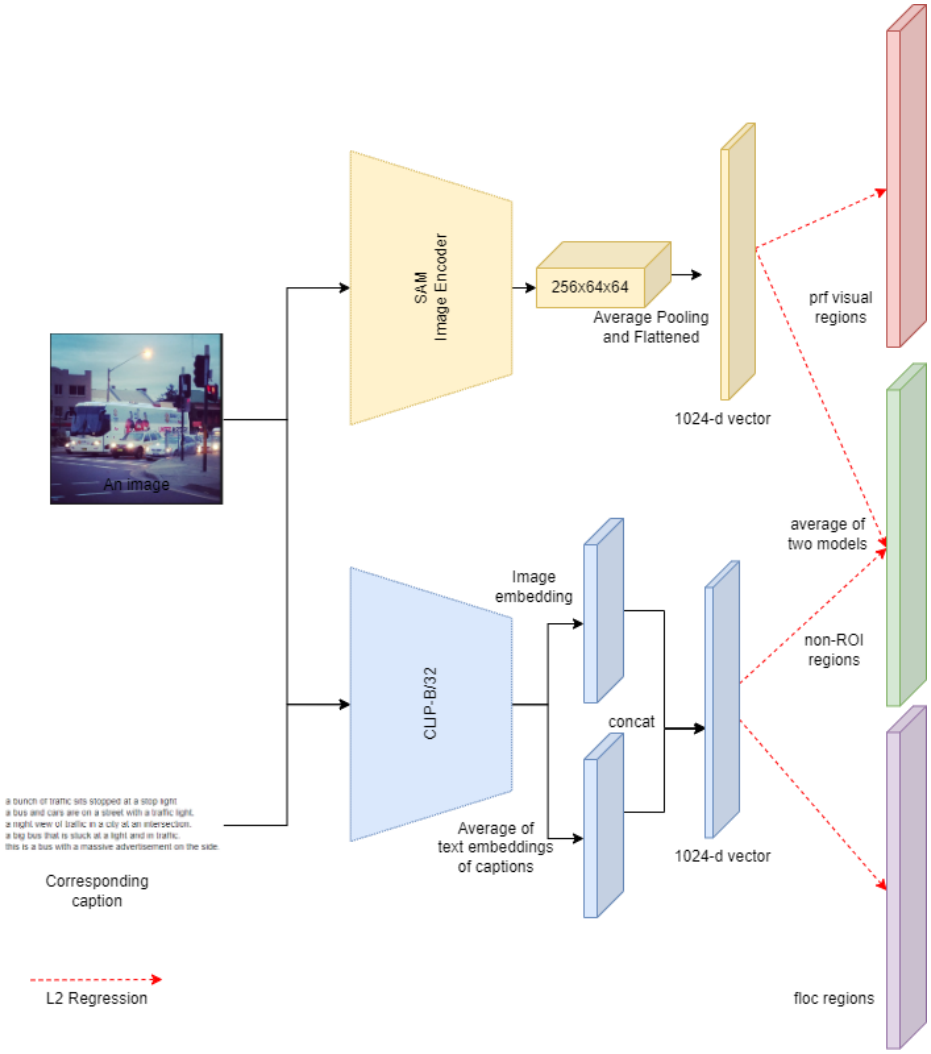

Fig. 2. Scheme of per ROI prediction

3 Conclusion

In our work, we utilized various pre-trained models to find the best visual representation of the natural scenes. We showed that the features from different models may exhibit higher correlations in different regions of the brain depending on their training objectives. Our best approach is a linear model mapping CLIP and SAM features to the fMRI responses. We hope that our work can contribute to the growing field of computational neuroscience. Our code is available on GitHub for anyone interested in replicating or building upon our methods.

References

1. Gifford A. T. et al. The algonauts project 2023 challenge: How the human brain makes sense of natural scenes //arXiv preprint arXiv:2301.03198. – 2023.
2. Allen EJ, St-Yves G, Wu Y, Breedlove JL, Prince JS, Dowdle LT, Nau M, Caron B, Pestilli F, Charest I, Hutchinson JB, Naselaris T, Kay K. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and computational intelligence. Nature Neuroscience, 25(1):116–126.
3. Lin, Tsung-Yi, et al. Microsoft COCO: Common Objects in Context. arXiv, 20 Feb. 2015. arXiv.org, https://doi.org/10.48550/arXiv.1405.0312.
4. Lin S., Sprague T., Singh A. K. Mind reader: Reconstructing complex images from brain activities Advances in Neural Information Processing Systems.
5. Scotti P. S. et al. Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors //arXiv preprint arXiv:2305.18274. – 2023.
6. High-resolution image reconstruction with latent diffusion models from human brain activity Yu Takagi, Shinji Nishimoto, bioRxiv 2022.11.18.517004; doi: https://doi.org/10.1101/2022.11.18.517004
7. Ozcelik F., VanRullen R. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion //arXiv preprint arXiv:2303.05334.–2023.
8. Radford A. et al. Learning transferable visual models from natural language supervision //International conference on machine learning.PMLR,2021
9. Fang, Yuxin, et al. "Eva: Exploring the limits of masked visual representation learning at scale." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
10. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
11. Kirillov, Alexander, et al. "Segment anything." arXiv preprint arXiv:2304.02643 (2023).