Brain Decodes Deep Nets

Huzheng Yang (huze@seas.upenn.edu)
University of Pennsylvania

James Gee* (gee@upenn.edu)
University of Pennsylvania

Jianbo Shi* (jshi@seas.upenn.edu)
University of Pennsylvania

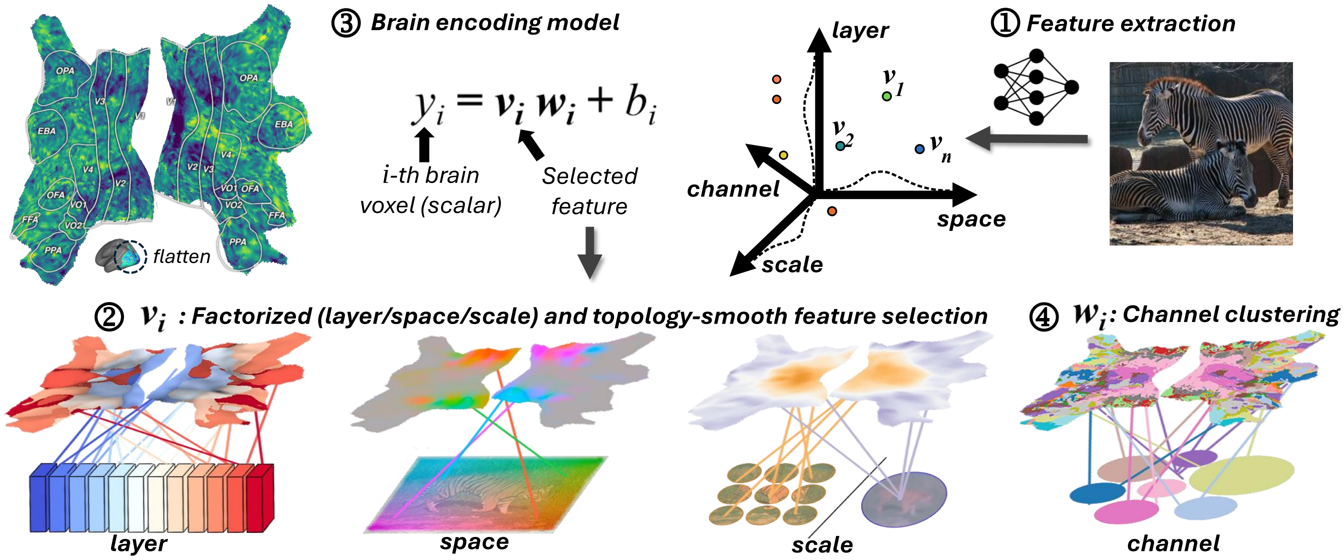

Figure 1: **Visualize Deep Networks in the Brain**. The training objective of the brain encoding model is to predict the brain's fMRI signal in response to an image stimulus. 3D visual brain surface is flattened into 2D for better visualization. ① Image features are extracted from a pre-trained network. ② Feature selection for each voxel is randomly initialized and learned using the brain encoding training objective. The selection is ***factorized*** in the layer/space/scale axis; the ***topological constraint*** improves selection smoothness and confidence. ③ Linearized brain encoding model. ④ After training, linear weights are used to cluster channels. We use the resulting brain-to-network mapping together with the known knowledge of the brain to answer the question *"how do deep networks work?"*.

Abstract

We developed a tool for visualizing and analyzing large pre-trained vision models by mapping them onto the brain, thus exposing their hidden inside. Our innovation arises from a surprising usage of brain encoding: predicting brain fMRI measurements in response to images. We report two findings. First, explicit mapping between the brain and deep-network features across dimensions of space, layers, scales, and channels is crucial. This mapping method, FactorTopy, is plug-and-play for any deep-network; with it, one can paint a picture of the network onto the brain (literally!). Second, our visualization shows how different training methods matter: they lead to remarkable differences in hierarchical organization of networks' intermediate layers. It also provides insight into fine-tuning: how pre-trained models change when adapting to small datasets. We found brain-like hierarchically organized network suffer less from catastrophic forgetting after fine-tuned.

Keywords: brain encoding; explainability; fine-tuning

Introduction

The brain is massive, and its enormous size hides within it a mystery: how it efficiently organizes many specialized modules with distributed representation and control. One clue it offers is its feed-forward hierarchical organization. This hierarchical structure facilitates efficient computation, continuous learning, and adaptation to dynamic tasks.

Deep networks are enormous, containing billions of parameters. Performances keep improving with more training data and larger size. It doesn't seem to matter if the network is trained under the supervision of labels, weakly supervised with image captions, or even self-supervised without human-provided guidance. Its sheer size also hides another mystery: as its size increases, it can be fine-tuned successfully to many unseen tasks.

*: Equal advising.

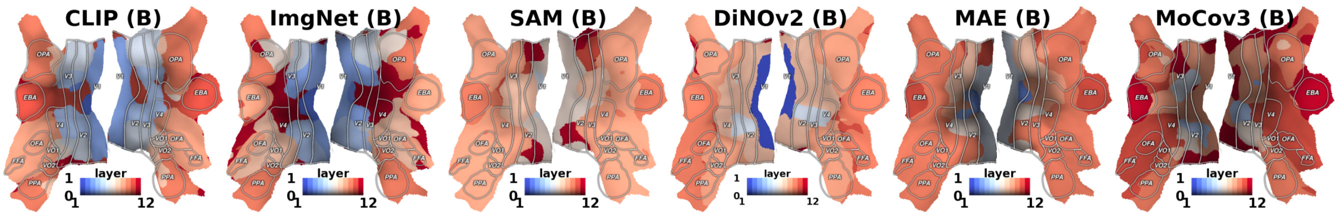

Figure 2: **Layer Selectors, Brain-Network Alignment**. All models are ViT architecture, 12 layers. Voxels colored by `argmax` of layer selection weight, brightness is confidence measurement, lower brightness means a *soft* selection of multiple layers.

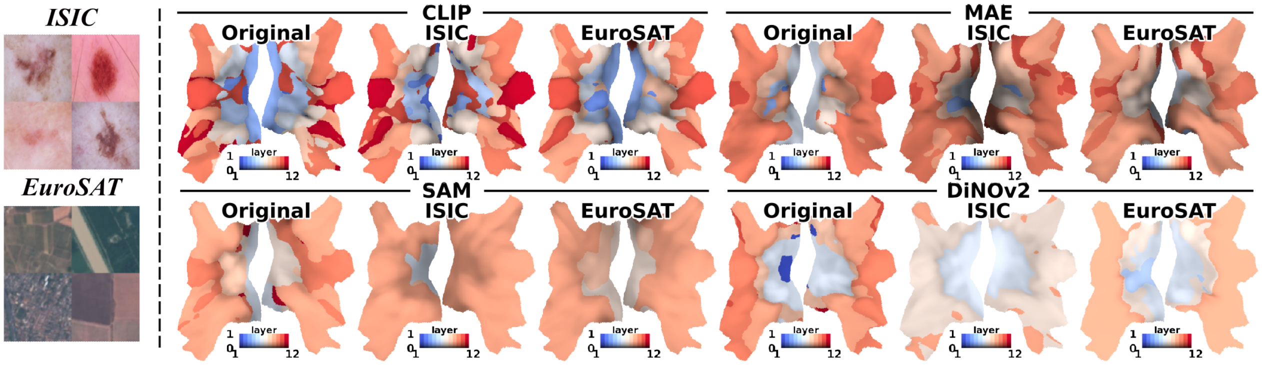

Figure 3: **Fine-tuned to Small Datasets**. ***Left***: example images from target datasets (*ISIC* and *EuroSAT*). ***Right***: layer selector before and after fine-tuning. The whole network is fine-tuned, then freeze to fit a new brain encoding model and layer selection. ***Insights***: CLIP and MAE fine-tune with less change in the existing computation.

What can these two massive systems, the brain and deep network, tell about each other? By identifying 'what' deep features are most relevant for each brain voxel fMRI prediction, we can obtain a picture of deep-net features mapped onto a brain, as shown by the brain-to-network mapping in Figure 1.

The key insight is that deep networks trained with the same architecture, but different objectives and data, produce drastically different computation layouts of intermediate layers, even if they can produce similar brain encoding scores and other downstream task scores. In Figure 2, we found intermediate layers of CLIP align hierarchically to the visual brain. However, there are unexpected non-hierarchical bottom-up and top-down structure in supervised classification and segmentation-trained models. Moreover, for many models, when scaling up in parameters and training data, they tend to lose hierarchical alignment to the brain, except CLIP, which improved hierarchical alignment to the brain after scaling up.

Suppose the brain's hierarchical organization is a template for efficient, modular, and generalizable computation; an ideal computer vision model should align with the brain: the first layer of the deep network matches the early visual cortex, and the last layer best matches high-level regions. Our fine-tuning results show that networks with more hierarchy organization tend to (qualitatively, in Figure 3) maintain their hidden layers better after fine-tuning on small datasets, thus suffering less (quantitatively, in Table 1) from catastrophic forgetting. We conjecture that better alignment to the brain is one way to find a robust model that adapts to dynamic tasks and scales better with larger models and more data.

Data and Methods

Brain Encoding Dataset We used Nature Scenes Dataset (NSD) (Allen et al., 2022) for this study. Briefly, NSD provides 7T fMRI scan when watching COCO images, 8 subjects recorded 40 hours each. After pre-processing and denoising (Kay et al., 2013; Prince et al., 2022), the brain encoding prediction target $\boldsymbol{Y} \in \mathbb{R}^{N\times 1}$ is beta weights (amplitude) of hemodynamic response (pulse) function, N denotes number of flattened voxels (vertices) in brain surface.

Table 1: Brain score dropped after fine-tuning. CLIP and MAE suffer less from catastrophic forgetting.

	Brain Score $R^2 \uparrow$		
Model / Fine-tune dataset	**Original**	**ISIC**	**EuroSAT**
CLIP	0.131	0.115	0.112
MAE	0.128	0.117	0.113
SAM	0.111	0.086	0.087
DiNOv2	0.128	0.085	0.082

Brain Encoding Model Figure 1 presents an overview of our methods. A frozen deep-net is used to extract image features, and linearized brain encoding model (Naselaris et al., 2011; Gifford et al., 2023) is trained to predict each brain voxel's response.

In this work, the key component is how each voxel selects input features $\boldsymbol{v}_i \in \mathbb{R}^{1\times C}$ for the linearized encoding model. The feature selection weights is the brain-to-network mapping showed in Figure 2 and 3.

Our fundamental innovations are two-fold. First, we propose a ***factorized*** feature selection across three independent dimensions (layer/space/scale). In the layer-axis, each voxel learned a random initialized and softmax-ed weight vector $\omega_i^{layer} \in \mathbb{R}^L$, where L is number of ViT layers. The layer selector weighted summed features from all layers. Factorized feature selection leads to a more robust and data-efficient estimation. Second, we enforce ***topology smoothness*** constraints: physically close-by brain voxels are forced to have similar layer selector weights. In practice, we use a learned MLP that take voxels' physical coordinates as input and output the selection weights (Lurz et al., 2021). The local smoothness constraints significantly reduce uncertainties in network-to-brain mapping.

Acknowledgement This work is supported by the funds provided by the National Science Foundation and by DoD OUSD (R&E) under Cooperative Agreement PHY-2229929 (The NSF AI Institute for Artificial and Natural Intelligence). Huzheng Yang and James Gee are supported in part by R01-HL133889, R01-EB031722, and RF1-MH124605.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., . . . Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*.

Gifford, A. T., Lahner, B., Saba-Sadiya, S., Vilas, M. G., Lascelles, A., Oliva, A., . . . Cichy, R. M. (2023). *The Algonauts Project 2023 Challenge: How the Human Brain Makes Sense of Natural Scenes.* arXiv:2301.03198.

Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F., & Wandell, B. A. (2013). GLMdenoise: a fast, automated technique for denoising task-based fMRI data. *Frontiers in Neuroscience*.

Lurz, K.-K., Bashiri, M., Willeke, K., Jagadish, A., Wang, E., Walker, E. Y., . . . Sinz, F. H. (2021). Generalization in data-driven models of primary visual cortex. In *International conference on learning representations.*

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*.

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022). Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife*.