Visual Feature-Based Brain Decoding Yields Weight Maps Better Aligned with Scene Understanding than Classification

Chenqian Le[*] **(cl6707@nyu.edu)**
Department of Electrical and Computer Engineering, New York University
370 Jay Street, Brooklyn, NY 11201

Nikasadat Emami[*] **(ne2213@nyu.edu)**
Department of Electrical and Computer Engineering, New York University
370 Jay Street, Brooklyn, NY 11201

Xujin Chris Liu (xl3942@nyu.edu)
Department of Electrical and Computer Engineering, New York University
370 Jay Street, Brooklyn, NY 11201

Xupeng Chen (xc1490@nyu.edu)
Department of Electrical and Computer Engineering, New York University
370 Jay Street, Brooklyn, NY 11201

Yao Wang[†] **(yaowang@nyu.edu)**
Department of Electrical and Computer Engineering, New York University
370 Jay Street, Brooklyn, NY 11201

∗ These authors contributed equally
† Corresponding author

Abstract

We introduce a brain decoding method for analyzing functional responses to visual perception using the Natural Scenes Dataset (NSD), where we use visual features of images from deep neural networks as a decoding target. Our method gives consistent results across various feature extraction methods and subjects. Using the resulting weight map in a follow-up classification task, our method achieves similar classification accuracy as a directly trained classifier yet offers broader applicability since no classification labels are needed. We show that our resulting weight maps are more closely aligned with the underlying task of human subjects compared to weight maps derived from classification-based decoding. The flexibility makes our method suitable for diverse decoding-style analysis with complex stimuli, where manual labeling might bias the results.

Keywords: Deep neural network; Brain visual decoding; fMRI

Introduction

Brain decoding is an important technique for deriving insights into the brain's functions by finding how voxel-level activation data can be used to predict certain stimuli or response variables[1, 2]. In this work, we investigate a simple tweak to the traditional classification-based decoding method: instead of using predefined classification labels [3, 4, 5, 6, 7], we propose to use pre-trained representations from deep neural network (DNN) models. In the interest of space, we focus on the visual cortex's response to natural scenes from the Natural Scenes Dataset (NSD), but our method can be easily applied to other domains, such as auditory processing. While some previous works have delved into regression-based approaches [8, 9, 10, 11, 12, 13], our methodology introduces a distinctive perspective. Our proposed new decoding method removes the need for existing stimuli labels and provides a weight map that better aligns with the underlying scene recognition process compared to classification-based decoding. Through a post-hoc classification test of scene classification, we show that our method preserves the class-related information even when not explicitly optimized for it, achieving a very similar performance as classification-based decoding. These advantages make our method a simple drop-in replacement for many decoding-style analyses involving complex responses or stimuli.

Methods

Dataset

Our research employs the Natural Scenes Dataset (NSD)[14], a comprehensive fMRI dataset captured at 7T featuring whole-brain, high-resolution measurements from eight healthy adults. Participants were exposed to thousands of color natural scenes from the extensively annotated Microsoft Common Objects in Context (COCO)[15] images during 30–40 scan sessions. We focus on data from four subjects who viewed identical stimuli, ensuring consistency in our analysis. This dataset is instrumental for investigating brain visual perception and pattern recognition.

Framework

Our framework integrates advanced feature extraction and dimension reduction techniques to analyze complex visual stimuli. We use pre-trained models, ResNet-50[16] and DINOv2[17], to extract visual features of a scene, followed by PCA[17] and UMAP[18], respectively, to reduce the dimension to two. For decoding analysis, we use the Nilearn[19]'s implementation of SpaceNet Decoder with Graph-Net regularization[20] to create both classification and regression weight maps. This methodology aids in producing interpretable brain weight maps. The overall pipeline is illustrated in Figure 1.

Post-hoc classification test

To quantify the informativeness of the resulting weight map from both methods, we use a post-hoc test to evaluate how well class-related information is preserved in the weight maps obtained from decoding analyses. Once we combine a final weight map from a decoding analysis by taking the max magnitude across all the sub-weight maps, regardless of the decoding target, we use it as a selection mask and decode the scene class from the FMRI analysis again. If the class information is preserved well in the first decoding step, the second post-hoc classification evaluation will yield high prediction accuracy. We evaluate at a number of different sparsity levels by thresholding the resulting weight maps at different levels.

Results

Post-hoc classification test

In our post-hoc classification evaluation of the weight maps, our label-free brain decoding method produces very similar levels of F1 score compared to traditional classification-based decoding, shown in Figure 2.b. Note that for a fair comparison, we selected an equivalent number of voxels from both the regression-based and classification-based methods. This indicates that the class-related information is adequately preserved in our method, even though this is not explicitly optimized for classification.

Analysis of the Visual Cortex Regions

In general, our method produces a similar weight distribution as classification-based decoding. As we can see from figure 2.a, we investigate different sub-regions of the visual part, including V1 to V5 cortex according to the Juelich atlas, and the following regions according to the Harvard-Oxford atlas: LG (Lingual Gyrus), LO-1 (Lateral Occipital Cortex superior division), LO-2 (Lateral Occipital Cortex inferior division), IC (Intracalcarine Cortex), CC (Cuneal Cortex), TOF (Temporal Occipital Fusiform Cortex), OFG (Occipital Fusiform Gyrus), and OP (Occipital Pole). Figure 2.a shows a high consistency across all combinations of DNNs and dimensionality reduction methods, with a higher average weight in areas associated with higher-level visual processing. While the findings from the classification-based decoding largely corroborate our results, disagreements appear in the V5 cortex and the LO-2 region, both associated with higher-order visual functions [21, 22].

Figure 2.a shows the mean correlation between voxel weight maps of different subjects. It is consistent across different subjects but decreases from 0.9 to 0.5 from V1 to V5. This trend might indicate a divergence in how subjects interpret combinations of high-level visual features but share similar processing of low-level visual features.

The visualization in Figure 2.c highlights significant weights in the Parahippocampal Place Area (PPA), a region integral to scene recognition and spatial memory, by the proposed approach.

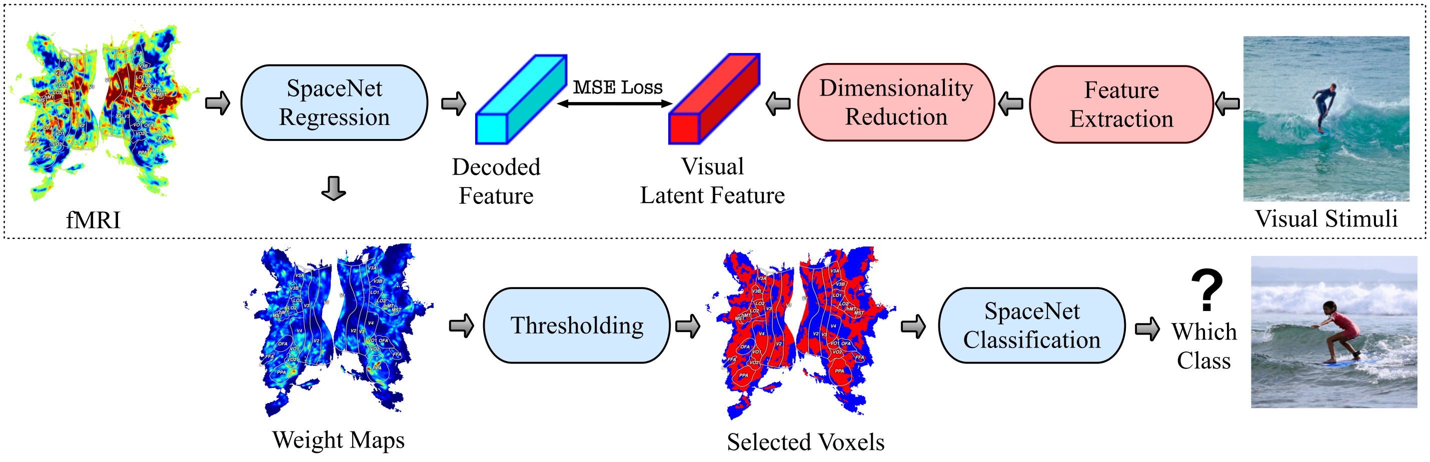

Figure 1: Pipeline Overview: Initially, visual stimuli are processed using a pre-trained deep neural network (either ResNet or DINOv2) to extract latent embeddings. These embeddings then undergo dimensionality reduction via PCA or UMAP to isolate fine-grained features. A linear regression model with Graph-Net regularization (SpaceNet) regresses these visual latent features. Subsequently, voxels of significant weights are selected for evaluation in an image classification task via thresholding.

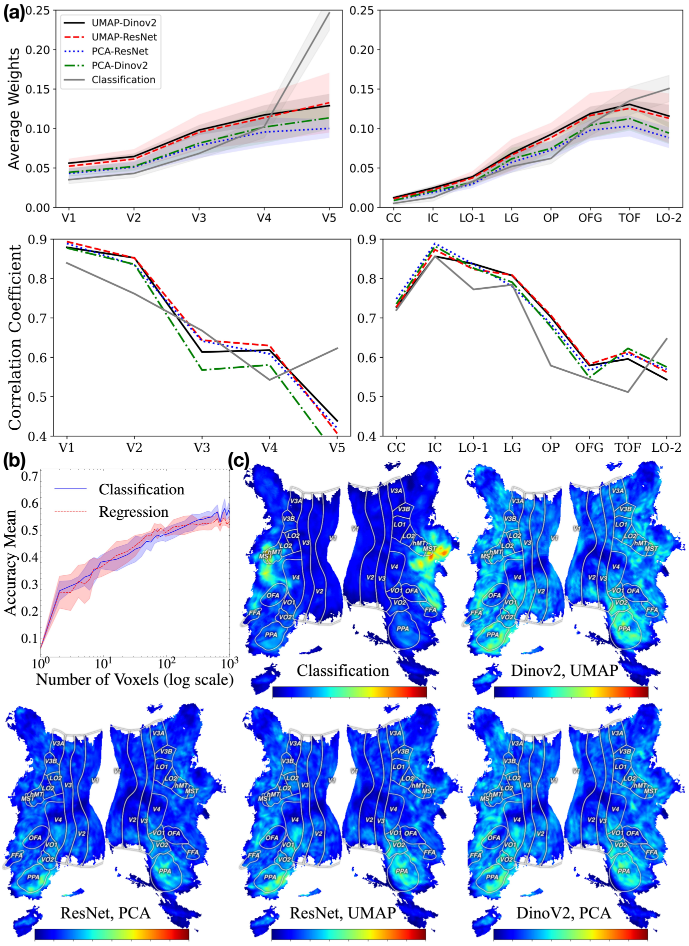

Figure 2: (a). Average voxel weights and the mean of weight correlation coefficients across subjects for visual subregions. (b). Image Classification Accuracy (c). Comparative Analysis of Weight Maps Across Methods: Average normalized values from the weight maps of each method across all subjects.

Higher weights are also prominent in the VO1 (visual occipital 1) and VO2 (visual occipital 2) regions, known for their roles in color recognition [23]. Conversely, the classification-based method assigns heavier weights to the Medial Temporal area, emphasizing motion perception[24]. We note that the underlying task of these fMRI scans is scene recollection, where participants recall previously viewed stimuli. This difference suggests that the weight maps produced by our method are better aligned with the underlying task of scene recognition.

Further analysis of weight progression from visual areas V1 to V5 shows an increase in weight intensity from basic visual processing in V1 to complex integrations in V5. This is particularly evident in classification-based methods, notably in the hMT(human Middle Temporal)/MST(Medial Superior Temporal) area known for motion sensitivity. This pattern highlights different neural engagements based on the decoding strategy, illustrating how these methods process visual information differently. This research enhances our understanding of the visual cortex's functional architecture and demonstrates the potential of advanced decoding techniques to reflect cognitive tasks in visual processing more accurately.

Conclusion and Discussion

We introduce a novel label-free brain decoding methodology using the Natural Scenes Dataset (NSD), where we replace the commonly used classification targets with features from pretrained deep neural networks, which removes the need for predefined classes or labels[25]. We demonstrated that this approach yields weight maps as informative as the traditional classification-based methods. A comparison of the weight maps shows that the regression-based method assigns weights in a way that better captures the underlying task of scene recognition, notably in brain regions like the Parahippocampal Place Area (PPA). Our proposed method provides a decoding analysis method that preserves relevant visual information, is consistent across parameter choices, and removes the reliance on hand-designed labels.

References

Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors, October 2023. arXiv:2305.18274 [cs, q-bio].

Huzheng Yang, James Gee, and Jianbo Shi. Brain decodes deep nets, 2024.

James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

Thomas A Carlson, Paul Schrater, and Sheng He. Patterns of activity in the categorical representations of objects. *Journal of cognitive neuroscience*, 15(5):704–717, 2003.

David D Cox and Robert L Savoy. Functional magnetic resonance imaging (fmri)"brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex. *Neuroimage*, 19(2):261–270, 2003.

John-Dylan Haynes and Geraint Rees. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature neuroscience*, 8(5):686–691, 2005.

Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–685, 2005.

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019.

Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstructing natural scenes from fmri patterns using bigbigan. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

Furkan Ozcelik, Bhavin Choksi, Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.

Y Liu, Y Ma, W Zhou, G Zhu, and N Zheng. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding. arxiv 2023. *arXiv preprint arXiv:2302.12971*.

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, February 2015. arXiv:1405.0312 [cs].

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. arXiv:1512.03385 [cs].

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. arXiv:2304.07193 [cs].

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. arXiv:1802.03426 [cs, stat].

Nilearn contributors. nilearn.

Logan Grosenick, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E. Taylor. Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage*, 72:304–321, May 2013.

Semir Zeki. Functional organization of a visual area in the posterior bank of the superior temporal sulcus of the rhesus monkey. *Journal of Physiology*, 236(3):549–573, 1974.

Rafael Malach, John B Reppas, Richard R Benson, Kenneth K Kwong, Hui Jiang, William A Kennedy, Patrick J Ledden, Thomas J Brady, Bruce R Rosen, and Roger BH Tootell. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, 92(18):8135–8139, 1995.

Steffie N. Tomson, Manjari Narayan, Genevera I. Allen, and David M. Eagleman. Neural networks of colored sequence synesthesia. *Journal of Neuroscience*, 33(35):14098–14106, 2013.

Julie Blumberg and Gabriel Kreiman. How cortical neurons help us see: visual recognition in the human brain. *The Journal of Clinical Investigation*, 120(9):3054–3063, 9 2010.

Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cerebral Cortex*, 28(12):4136–4160, 10 2017.