DOI: 10.3217/978-3-99161-050-2-084

Do Complex Vision Models Improve Feature Alignment with fMRI for Neural Decoding of Visual Stimuli?

M. Finocchiaro[1*], S. Calcagno[1], F. Proietto Salanitri[1], L. Passarrello[1], C. Spampinato[1]

[1]*PeRCeiVe Lab, University of Catania, Italy*

*E-mail: finocchiaro.marco@phd.unict.it

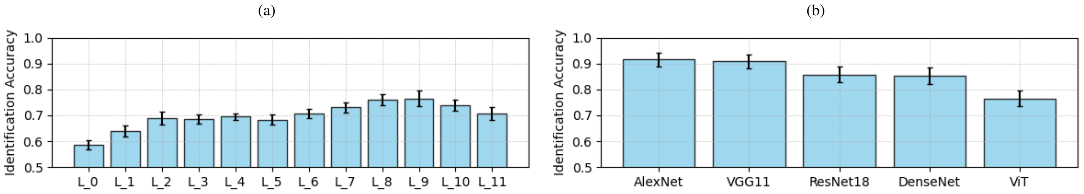

Figure 1: **(A)** Performance of ViT self-attention layers in aligning with fMRI data, highlighting the impact of layer depth on model identification accuracy. **(B)** A schematic comparison showing that more complex networks, such as ViT, are outperformed by simpler CNN architectures.

Introduction: Recent advancements in computer vision, particularly with Vision Transformers (ViTs) and foundation models, have led to significant improvements across various tasks by utilizing self-attention mechanisms to capture intricate patterns and global relationships in visual data. This study explores whether these advanced models, compared to traditional convolutional neural networks (CNNs), offer better alignment with neural data, particularly fMRI responses. We specifically investigate the impact of model complexity on interpreting brain activity by evaluating the alignment between fMRI data and features from several neural network architectures, including ViT, ResNet18, DenseNet [1, 2, 4], and simpler models like VGG11, and AlexNet [5, 7].

Materials, Methods, and Results: We used the fMRI dataset in [3], which links visual stimuli from ImageNet categories to corresponding fMRI data. The training set consisted of 1200 samples from 150 image classes (8 images per class), while the test set included 1750 recordings from 50 image classes, each observed 35 times. Preprocessed data through motion correction, voxel normalization, and ROI selection, was used for training AlexNet, VGG11, ResNet18, DenseNet, and ViT.

Performance in terms of identification accuracy [3] across features extracted from different layers, shows that simpler models, like VGG11 and AlexNet, outperform more complex architectures such as ResNet, DenseNet, and ViT. Notably, ViT exhibited the lowest performance, highlighting its reduced compatibility with fMRI signals.

With ViT models increasingly explored for decoding brain activity due to their representational power [6], we also investigated their alignment with fMRI data across all layers. As shown in Fig.1a, ViT exhibits a distinct trend compared to CNNs, where performance typically improves in early layers, peaks at intermediate layers, and declines in deeper layers[3]. In ViT, alignment gradually improves from early to deeper layers, with a performance drop in the final layer. This discrepancy may arise from fundamental differences in processing: CNNs hierarchically encode perceptual features in early layers and abstract semantics in deeper ones, while ViTs use self-attention to refine token relationships, emphasizing global context and abstraction throughout the network, especially in the last layer. Thus, the lower performance by ViT likely suggests that its highly abstract and complex representations are less compatible with the predominantly low- and mid-level features captured by fMRI signals.

Conclusion: This work indicates that higher architectural complexity does not always translate to improved compatibility with biological signals, underscoring the importance of considering model complexity and representation characteristics when applying machine learning to decode brain activity.

Acknowledgments and Disclosures: M. Finocchiaro is a PhD student enrolled in the National PhD in Artificial Intelligence - ”Health and Life Sciences”, Università Campus Bio-Medico di Roma.

References:

[1] A. Dosovitskiy. “An image is worth 16x16 words: Transformers for image recognition at scale”. *arXiv preprint* (2020).

[2] K. He et al. “Deep residual learning for image recognition”. *CVPR*. 2016.

[3] T. Horikawa and Y. Kamitani. “Generic decoding of seen and imagined objects using hierarchical visual features”. *Nat. Comm.* (2017).

[4] G. Huang et al. “Densely connected convolutional networks”. *CVPR*. 2017.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet classification with deep convolutional neural networks”. *NeurIPS* (2012).

[6] Ken Shirakawa et al. “Spurious reconstruction from brain activity: The thin line between reconstruction, classification, and hallucination”. *Journal of Vision* 24.10 (2024), p. 321. DOI: 10.1167/jov.24.10.321.

[7] K. Simonyan. “Very deep convolutional networks for large-scale image recognition”. *arXiv preprint* (2014).