Can Scene Graph Properties Explain the Neural Encoding Performance of Vision Transformers?

Helena Balabin, Rik Vandenberghe, Marie-Francine Moens
KU Leuven, Leuven, Belgium

Abstract

Neural encoding models allow for the exploration of hypotheses about cognitive processes by linking brain activations to representations derived from large language or image models. However, such representations often remain poorly understood, limiting the interpretability of neural encoding models. Therefore, we set out to examine the effect of scene graph properties on image model representations and neural encoding performances to functional magnetic resonance imaging (fMRI) data from the Natural Scenes Dataset (NSD). Specifically, we used the overlap between the NSD and the Visual Genome to characterize each image using the number of relationships, objects and depth of the accompanying scene graph annotations. We found that relationships and depth measures could be decoded more accurately both from fMRI activations and from image embeddings compared to objects, aligning with an afforance-based scene perception approach.

Keywords: neural encoding; multivariate pattern analysis; scene graphs; vision transformer

Introduction

Neural encoding refers to the prediction of neural representations, such as activations obtained from functional magnetic resonance imaging (fMRI), from stimuli like images or sentences (Naselaris, Kay, Nishimoto, & Gallant, 2011). More recently, large language or vision models have been leveraged to represent the stimuli, allowing for the exploration of comprehensive hypotheses about both cognitive processes and the inner workings of foundation models through neural encoding (Oota et al., 2024).

When comparing the layer-wise alignment of language models with neural activations, previous research suggests that middle layers result in the best neural encoding performance (Jain & Huth, 2018; Toneva & Wehbe, 2019). This alignment is substantially influenced by syntactic information implicitly encoded in language model representations, in particular the constituency parse tree depth (Oota, Gupta, & Toneva, 2023). In vision models, however, it remains unclear how the structure of a natural scene may influence the alignment of a vision model to brain activations, forming a critical gap for testing more targeted hypothesis about cognitive and model mechanisms.

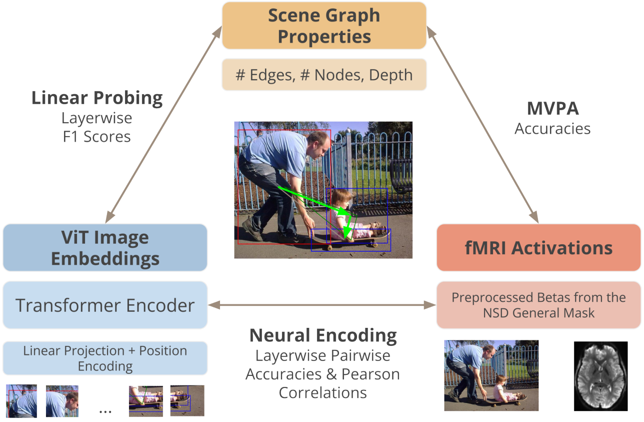


Figure 1: Overview of our proposed three-way comparison of ViT embeddings, fMRI activations and scene graph properties. *fMRI, functional magnetic resonance imaging; MVPA, multivariate pattern analysis; NSD, natural scenes dataset; ViT, visual transformer*

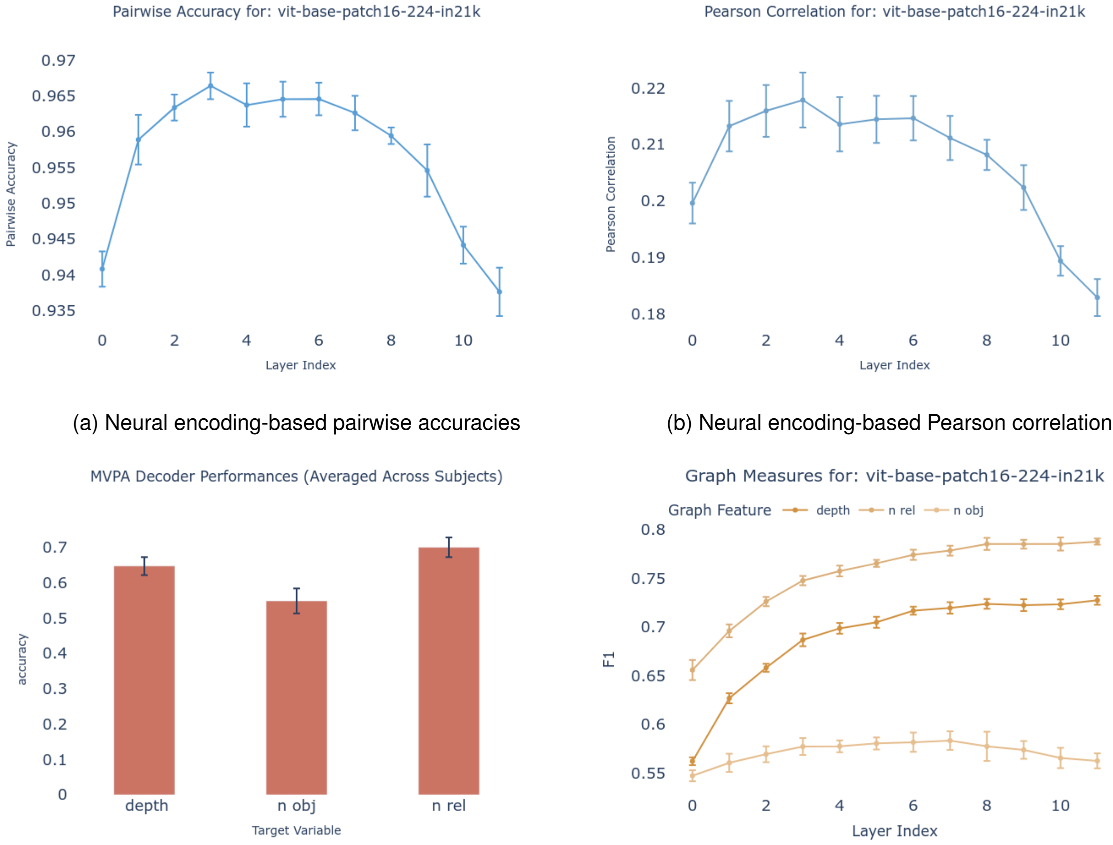
Therefore, this study investigates the effect of scene graph properties on the neural encoding performance between the visual transformer (ViT) model (Dosovitskiy et al., 2021) and fMRI activations from the Natural Scenes Dataset (NSD) (Allen et al., 2022) using scene graph annotations from the stimuli that overlap with the Visual Genome (VG) (Krishna et al., 2017). We hypothesize that relationships, and specifically relationship-based depth measures are decodable from both neural data and image embeddings.

Methods

Data To investigate the link between scene graph properties and neural encoding performances, we used the 73,000 NSD (Allen et al., 2022) images as a starting point, and focused on a subset of 28,459 images that had accompanying scene graph annotations from the VG (Krishna et al., 2017).

Scene Graph Properties For each image, a scene graph was constructed from the VG metadata. To ensure high quality scene graphs, we only included relationships that contained at least one living being as a subject or object. Next, we derived three characteristics from each graph: (1) The number of relationships/edges, (2) number of objects/nodes and (3) depth of the graph (i.e., the longest shortest distance between any two nodes in the graph, as the graphs are not necessarily trees). Then, we binarized each characteristic as either low or high based on whether it was below or above the median value across all graphs.

Neural Encoding We first determined the layerwise neural encoding performances of the ViT model and fMRI activations from the regions of interest specified by the NSD general mask (see lower part of Figure 1). Performances were separately



(a) Neural encoding-based pairwise accuracies

(b) Neural encoding-based Pearson correlations

(c) MVPA accuracies

(d) Linear probing-based F1 scores

Figure 2: Comparison of visual transformer-based image embeddings, fMRI activations and scene graph properties using neural encoding, MVPA and linear probing. *fMRI, functional magnetic resonance imaging; MVPA, multivariate pattern analysis*

evaluated using pairwise accuracies and Pearson correlations (Oota et al., 2024) for the embeddings of each ViT layer.

Multivariate Pattern Analysis We then used multivariate pattern analysis (MVPA) to investigate the accuracy with which the three scene graph characteristics can be predicted from the fMRI activations (see right part of Figure 1).

Linear Probing Finally, and analogously to the MVPA, we examined whether the scene graph properties can be predicted from the ViT embeddings using linear probing for each layer (see left part of Figure 1). We evaluated the linear probing performance using F1 scores.

Results

The neural encoding accuracies & Pearson correlations, MVPA and linear probing results are shown in Figure 2a & 2b, Figure 2c and Figure 2d, respectively. Both layerwise neural encoding pairwise accuracies and pearson correlations exhibited a similar pattern, with the middle layers yielding the highest performances, and substantially lower performances for the first and last layers. Further, we observed the highest MVPA accuracies for the number of relationships, closely followed by the depth. In a similar manner, the linear probing results revealed that that across all layers, the prediction of the number of objects barely surpassed chance performance. Conversely, the number of relationships and depth measures led to considerably higher linear probing-based F1 scores, especially across middle and late model layers.

Discussion

We aimed to investigate the effect of the structure of a scene on neural encoding performances between fMRI activations and ViT image embeddings. Our neural encoding results mirror layerwise performance differences that were observed in language models (Jain & Huth, 2018; Toneva & Wehbe, 2019). With regard to scene graph properties, we found that relationships and depth measures could be decoded more accurately both from fMRI activations and from ViT image embeddings compared to objects. This finding aligns with an affordance-based scene perception approach (Gibson, 1977), which states that the perception of a scene is primarily defined by its enabled actions. Lastly, the binarization of the scene graph properties presents a considerable reduction, and future work should explore more granular settings.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., . . . Kay, K. (2022, January). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. (Number: 1 Publisher: Nature Publishing Group)

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . Houlsby, N. (2021, June). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929*.

Gibson, J. J. (1977). The theory of affordances. In *Perceiving, acting and knowing: Towards an ecological psychology* (p. 67). Boston: Lawrence Erlbaum.

Jain, S., & Huth, A. G. (2018, December). Incorporating context into language encoding models for fMRI. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 6629–6638). Red Hook, NY, USA: Curran Associates Inc.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., . . . Fei-Fei, L. (2017, May). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, *123*(1), 32–73.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011, May). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410.

Oota, S., Gupta, M., Bapi, R. S., Jobard, G., Alexandre, F., & Hinaut, X. (2024, December). Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey). *Transactions on Machine Learning Research*.

Oota, S., Gupta, M., & Toneva, M. (2023, December). Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, *36*, 18001–18014.

Toneva, M., & Wehbe, L. (2019, November). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Proceedings of the 33rd Conference on Neural Information Processing Systems.* Vancouver, Canada.