Image Embeddings Informed by Natural Language Improve Predictions and Understanding of Human Higher-level Visual Cortex

Aria Y. Wang[1,2], Michael J. Tarr[1,2,3], Leila Wehbe[1,2]
[1]Neuroscience Institute [2]Machine Learning Department [3]Department of Psychology

Motivation

The way a model is trained affects the representations it learns...

ImageNet trained models:
"A baseball player"

Contrastive Language-Image Pre-training (CLIP) model:
"A batter has just attempted to hit the ball being pitched to him while waiting at home plate."

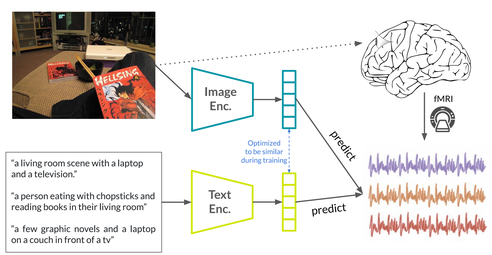
How do the different representations affect our ability to predict a brain area, and what does that tell us about how the brain interprets scenes?

Methods

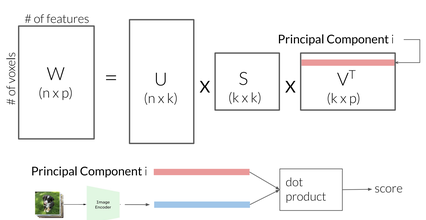
Natural Scene Dataset (NSD) – A 7T large scale fMRI dataset[1]

- 8 participants
- 9,000–10,000 distinct color natural scenes from COCO dataset
- task: long-term continuous image recall

Model pipeline for brain prediction

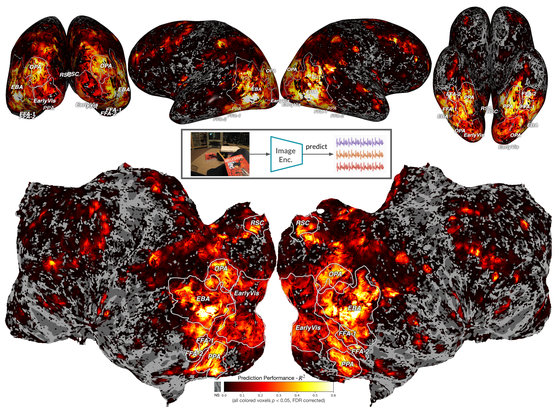


PCA and maximizing stimuli for PC

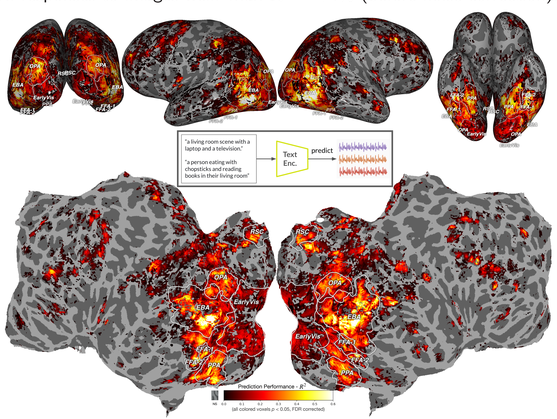


Model Performance

Representations from CLIP visual encoder predict fMRI responses to images extremely well. Max R^2 = 78% (before noise correction).

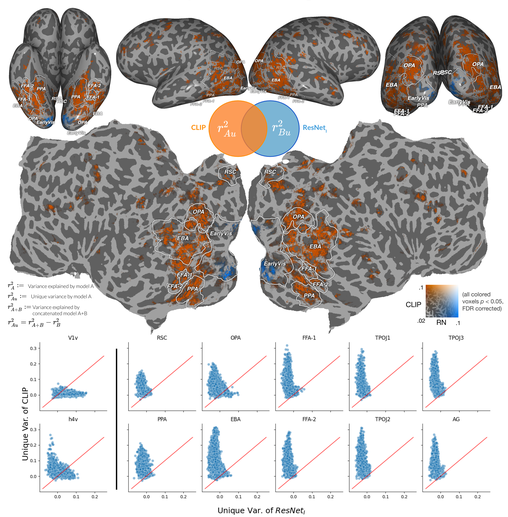


Representations from CLIP text encoder of image captions can also predict fMRI responses to images well. Max R^2 = 74% (before noise correction).



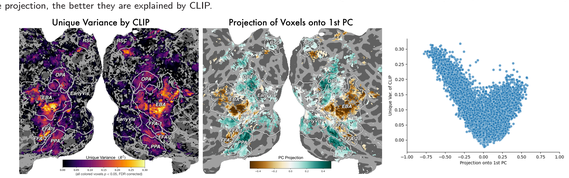
Unique Variance by CLIP

CLIP visual encoder with a ResNet50 backbone explains significantly more variances in various ROIs compared to a ImageNet trained ResNet50.



Representation of people scene may account for CLIP's unique variance.

Voxels that are explained most by CLIP **(Left)** overlap largely with the voxels that lie on the negative side when projected onto the 1st PC **(Middle)**. **(Right)** Voxel-wise scatter plot validates that for voxels that lie on the negative side of 1st PC projection, the further down they lie on the projection, the better they are explained by CLIP.

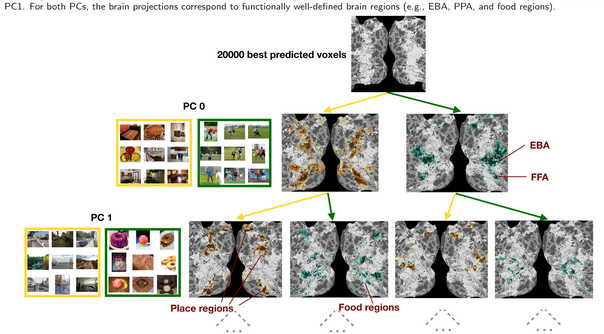


Left: Images that are most aligned with two ends of first PC are extracted (see method), and shown in "+" and "−". **Right**: Category distribution of these two groups of images validates that images on the negative side consist more people, animal, and sports, compared to images in the other group.

Visualizations of Semantic Dimension and Distances

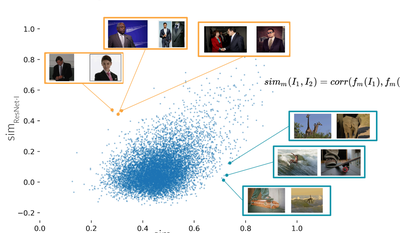
Learned prediction model with CLIP captures important semantic dimensions.

Top images are visualized for each PCs. Animate and inanimate images are separated by PC0, while scenes versus food images are separated by PC1. For both PCs, the brain projections correspond to functionally well-defined brain regions (e.g., EBA, PPA, and food regions).



CLIP and ResNet$_I$ represents images differently.

Pairwise similarities of the CLIP and ResNet$_I$ representations for 1000 randomly selected stimulus images picks out semantically and visually pairs, respectively.



Conclusions

Multimodal representation (e.g. from CLIP)

- gives better prediction across the high level visual cortex
- provides an effective way of mapping semantic information in the visual processing pathways
- allows for new ways of uncovering semantic basis of the brain

References

[1] Allen, E.J. et al., A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nature Neuroscience (2021).