Do Masked Autoencoders Learn a Human-Like Geometry of Neural Representation? Divergence and Convergence Across Brains and Machines During Naturalistic Vision

Dept. of Psychology and Neuroscience, Boston College

Hamed Karimi, Stefano Anzellotti

@hamedk72.bsky.social
hamedk72@gmail.com

SCCN

Categorization performance correlates with model-to-brain similarity in static vision models

Models with better categorization performance over static stimuli (i.e., images) are often reported to be better in capturing the variation of neural activity [1].

Masked Autoencoders (MAE) are remarkably effective at categorization. Making them a promising candidate model for neural responses [2].

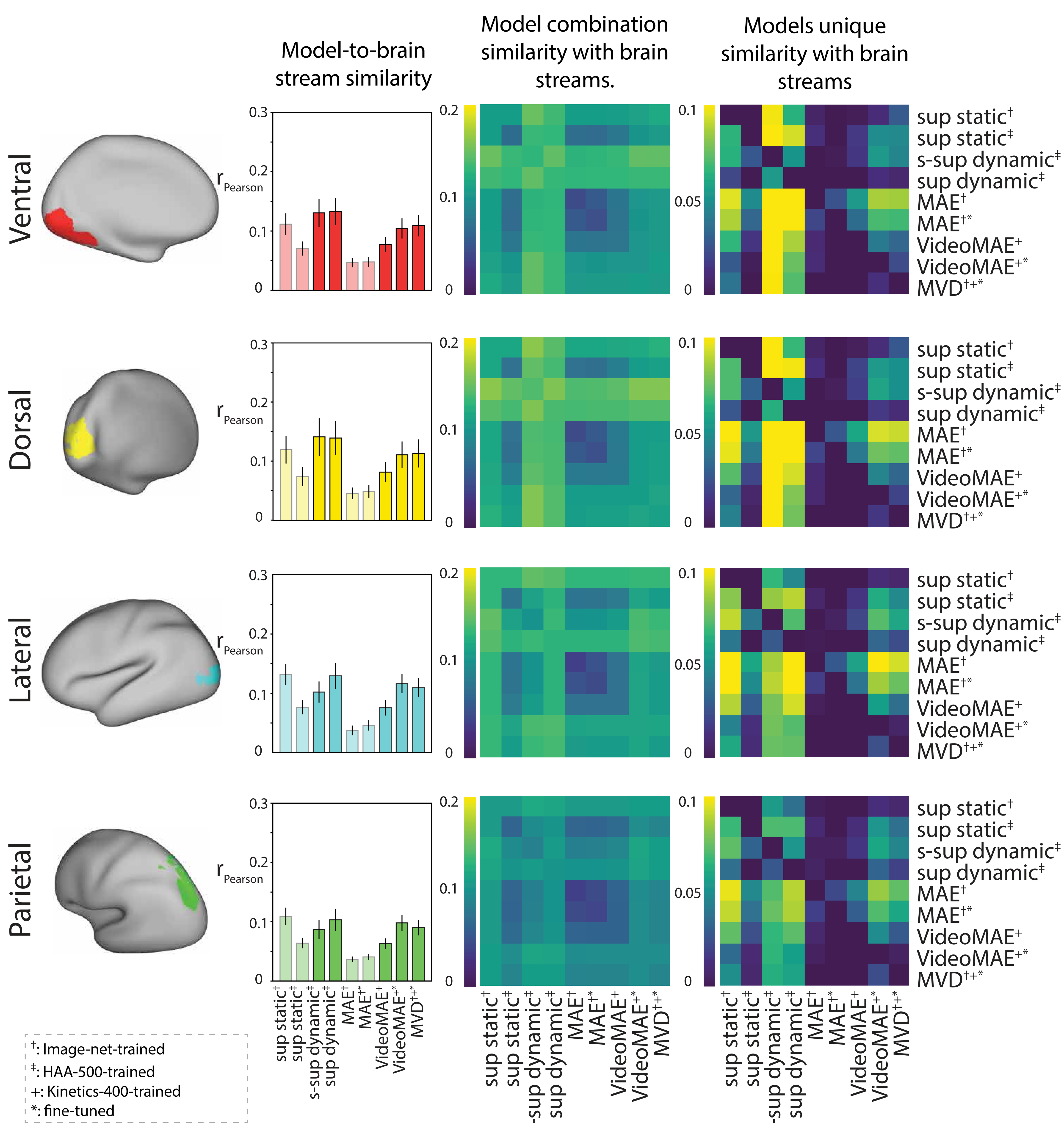
In addition to learning static spatial dependencies, Video-MAEs learn a representation of temporal information in dynamic stimuli (i.e., videos) [3].

In the current study, we asked how well MAE models can capture the variation of neural activity, and how they are different from CNN in terms of their similarity with different brain streams.

Dynamic and Static models of human brain

Model	Input	Output	Training Dataset	#Selected Layers
Supervised static	Image	Object identity	ImageNet	11
Supervised static	Image	Action identity	HAA-500	11
Self-supervised dynamic	Video	Optic flows	HAA-500	11
Supervised dynamic	Video	Action identity	HAA-500	11
Pre-trained Masked Autoencoder (MAE)	(Masked) image	(Unmasked) image	ImageNet	12
Fine-tuned Masked Autoencoder (MAE)	Image	Object identity	ImageNet	12
Pre-trained Video Masked Autoencoder	(Masked) video	(Unmasked) video	Kinetics-400	12
Fine-tuned Video Masked Autoencoder	Video	Action identity	Kinetics-400	12
Pre-trained Masked Video Distillation	(Masked) video	MAE and VideoMAE high-level features	Kinetics-400	12

Results



†: Image-net-trained
‡: HAA-500-trained
+: Kinetics-400-trained
*: fine-tuned

sup: supervised
s-sup: self-supervised
MAE: masked autoencoder
MVD: masked video distillation

Light color: static models
Bold color: dynamic models

element (*i*, *j*) displays the similarity (Pearson's r) betwee combined features of model *i*, and *j* with each brain stream

element (*i*, *j*) displays the similarity (Pearson's r) between each brain stream and model *j* features when controlled for model *i* features variation.

Conclusion

- Image MAE models show little correspondence with the neural activity in different brain streams.
- MAE models which represent dynamic information (VidoeMAEs and MVD) capture neural responses variation better but not as well as dynamic CNN models.
- Models based on optic flow representations (s-sup dynamic and sup dynamic) accounted for unique variance in all brain streams, even compared to Video-MAEs and MVD.
- Despite learning to represent large-scale spatial and temporal dependencies of stimuli that are effective for categorization, Image and Video MAE models rely on a mechanism that differs from the human brain.
- Future research should focus on identifying tasks that reveal the differences between MAEs and human performance, as categorization tasks, where MAEs excel, do not adequately highlight this divergence.

References

[1] Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences of the United States of America, 111(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

[2] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16000-16009).

[3] Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems, 35, 10078-10093.