Dynamic, social vision highlights gaps between deep learning and human behavior and neural responses

Kathy Garcia* (kgarci18@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N Charles St, Baltimore, MD, USA 21218

Emalie McMahon* (emaliemcmahon@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N Charles St, Baltimore, MD, USA 21218

Colin Conwell (cconwell2@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N Charles St, Baltimore, MD, USA 21218

Michael F. Bonner (mfbonner@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N Charles St, Baltimore, MD, USA 21218

Leyla Isik (lisik@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N Charles St, Baltimore, MD, USA 21218

*co-first authors

Abstract

To date, deep learning models trained for computer vision tasks are the best models of human vision. This work has largely focused on behavioral and neural responses to static images, but the visual world is highly dynamic, and recent work has suggested that in addition to the ventral visual stream specializing in static object recognition, there is a lateral visual stream that processes dynamic, social content. Here, we investigated the ability of 350+ modern image, video, and language models to predict human ratings of visual-social content of short video clips and neural responses to the same videos. We find that unlike prior benchmarks, even the best image-trained models do a poor job of explaining human behavioral judgements and neural responses. Language models outperform vision models in predicting behavior but are less effective at modeling neural responses. In early and mid-level lateral visual regions, video-trained models predicted neural responses far better than image-trained models. However, prediction by all models was overall lower in lateral than ventral visual regions of the brain, particularly in the superior temporal sulcus. Together, these results reveal a key gap in modern deep learning models' ability to match human responses to dynamic visual scenes.

Keywords: vision; social perception; action recognition; fMRI; deep learning; NeuroAI

Introduction

Human cognition is remarkably attuned to social visual cues, enabling us to quickly and automatically recognize and interpret interactions and intentions of others (McMahon & Isik, 2023). Recent work has suggested that social visual scenes are processed along the lateral surface of the brain originating in the early visual cortex (EVC) and extending to the superior temporal sulcus (STS) (Wurm, Caramazza, & Lingnau, 2017; Pitcher & Ungerleider, 2021). Despite the importance of dynamic, social perception, NeuroAI has focused almost exclusively on the match between humans and AI in static scenes (Allen et al., 2022; Kriegeskorte, 2015). To address this gap, we employ large-scale benchmarking of over 350 image, language, and video models, including state-of-the art (SOTA) models, to compare these models to human behavior and fMRI responses to naturalistic social videos (Figure 1).

Methods

Behavioral and neural data

Behavioral judgements and neural data were previously published in (McMahon, Bonner, & Isik, 2023) and are publicly available. Briefly, the authors collected behavioral ratings of the visual social scene on 250 3-second videos of dyadic social actions and showed these same videos to fMRI participants over many sessions, obtaining high-quality data for model evaluation. The dataset includes eight behavioral ratings collected on a likert scale (N $\geq$ 10 per rating per video): the spatial expanse of the scene (i.e., close up versus far away scenes), the extent to which the video contained an object-directed action, distance between people in the video, extent to which two people are facing, whether people are engaged in a joint action (like dancing or fighting), whether people are communicating, valence, and arousal.

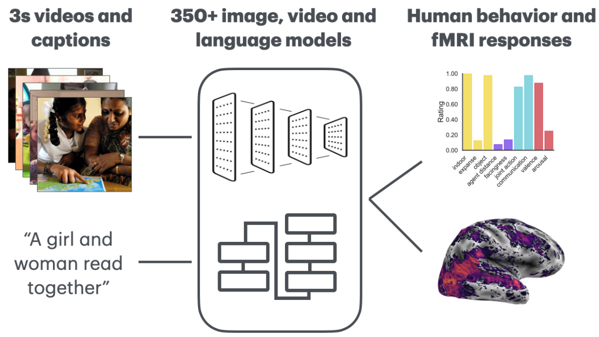


Figure 1: Features from image, video, and language models were extracted from either videos or captions and used to predict behavioral ratings and fMRI responses to the videos.

To evaluate the language models, we additionally had a new group of 150 online participants annotate the actions and interactions of the agents in the videos in a single sentence. We collected at least five unique captions for each video. Captions were cleaned by removing participants whose captions were determined to be 2.5 standard deviations away from the mean of other raters in the embedding space of all-MiniLM-L12-v1 implemented in Hugging Face (Wolf et al., 2020).

Model selection

We selected models with a broad range of architectures, training sets and objectives. We tested over 300 image models from collections including Torchvision and Pytorch-Image-Models libraries, VISSL, OpenAI's CLIP, and Dectectron2. We selected eight video models. Notable video models included Facebook's SlowFast and TimeSformer models. Image and video models were selected to represent a comprehensive cross-section of high-level visual tasks, and include convolutional and transformer architectures. Fifteen language models were selected based on performance in natural language processing tasks, focusing on sentence-transformers, including variants from CLIP and GPT-2 architectures. We tested fewer video and language models relative to image models due to their availability and computational costs, respectively. However, this only strengthens our conclusions when either model class outperforms image models.

Behavioral and neural alignment

For each model, we extracted the activations from every layer of the model and used optimized leave-one-out Ridge regression and 4-fold cross validation in the training set as implemented in (Conwell, Prince, Kay, Alvarez, & Konkle, 2023) to find the best fitting model layer. We then evaluated the best model layer on the test set. Performance was determined as the correlation between the predicted behavioral ratings or neural activation and true data.

To provide image models with information from across the video, each model received 7 evenly sampled frames from the video, which were then averaged in the model's feature space. Preliminary analyses showed very similar results when frames were concatenated. Similarly, activations were extracted for every caption from the language models and then averaged.

Results

Behavioral prediction

All models perform similarly well, near the level of human reliability, for spatial expanse, which is a static scene feature. For all other social action features, traditional image-trained models perform well below the level of human agreement (Figure 2). On average for most features, video models provide a slight performance boost over image models, and language models are more predictive than image and video models, particularly for the higher-level social features. Interestingly, the larger models with more training (e.g., GPT2, X3D, CLIP) did not always provide the best match to behavior.

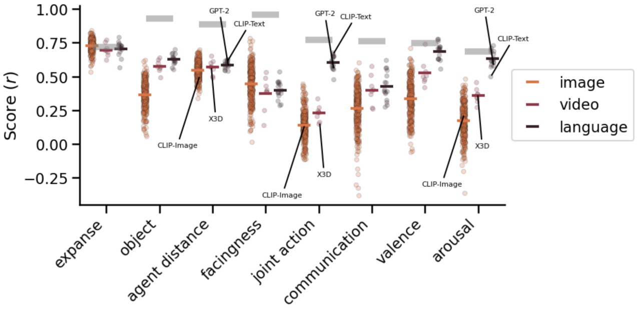


Figure 2: Performance of each model (dots) at predicting the human judgments of different visual-social features of the videos. The solid colored lines represent the mean of each class of models. The gray bars are the split-half reliability of ratings across participants. Select models are noted.

Neural prediction

We examined voxelwise neural responses averaged in several regions of interest (ROIs) in the lateral visual stream (EVC, middle temporal area, MT, extrastriate body area, EBA, lateral occipital cortex, LOC, and social-interaction selective regions in posterior STS, pSTS, and anterior STS, aSTS) and also included two ventral ROIs for comparison (fusiform face area, FFA, and parahippocampal place area, PPA). On average, video models outperform image and language models in all ROIs in the lateral stream. In particular, video models provide a large performance boost over image models in early and mid-level lateral visual regions such as MT and EBA. However, in more anterior regions along the STS, all models have lower predictive power, in contrast to ventral regions where image and video models are at or near the level of the noise ceiling (Figure 3). Similar trends can be seen in a whole brain analysis (Figure 4).

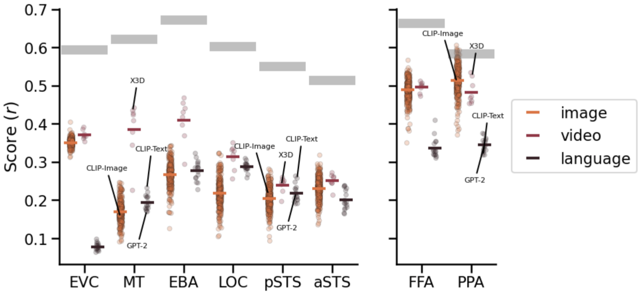
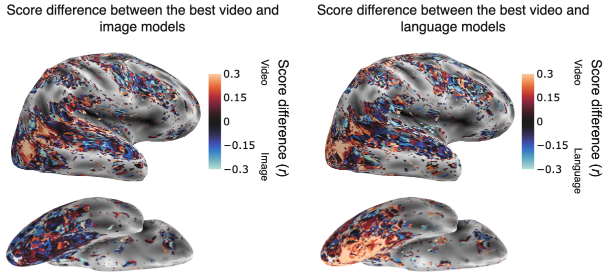


Figure 3: Performance of models at predicting neural responses in lateral (left) and ventral (right) ROIs averaged across participants. Gray bars are the trial-wise split-half reliability averaged across participants. Select models are noted.



Score difference between the best video and image models
Score difference between the best video and language models
0.3
0.15
0
−0.15
−0.3
Video
Image
Score difference (r)
0.3
0.15
0
−0.15
−0.3
Video
Language
Score difference (r)

Figure 4: The difference in performance between the best performing video and image model (left) and language model (right) in one representative subject.

Discussion

The results provide a comprehensive assessment of image, video, and language models, including SOTA models, in matching human responses to dynamic social perception. Despite their strong match to other areas of human visual behavior and brain responses (Geirhos et al., 2021; Conwell et al., 2023), the performance of image-trained models was quite poor in predicting human judgments of various visual-social features of short video clips and brain responses along the lateral stream. Language models and video models were better at predicting behavior and brain responses, respectively, perhaps because both can better capture rich event structure that image models cannot. However, no model could accurately match human responses across brain and behavior. Together, these results reveal a substantial gap in current AI models' ability to match human visual behavior to dynamic social scenes as has been recently identified for naturalistic face perception (Jiahui et al., 2023), and highlight the importance of studying vision dynamic, social contexts.

Acknowledgments

This work was funded in part by NSF GRFP DGE-1746891 awarded to E.M. and NIMH R01MH132826 awarded to L.I.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., . . . Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. doi: 10.1038/s41593-021-00962-x

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2023). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*. doi: 10.1101/2022.03.28.485868

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 23885–23899). Curran Associates, Inc.

Jiahui, G., Feilong, M., Visconti di Oleggio Castello, M., Nastase, S. A., Haxby, J. V., & Gobbini, M. I. (2023). Modeling naturalistic face processing in humans with deep convolutional neural networks. *Proceedings of the National Academy of Sciences*, *120*(43), e2304085120. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.2304085120

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446. doi: 10.1146/annurev-vision-082114-035447

McMahon, E., Bonner, M. F., & Isik, L. (2023). Hierarchical organization of social action features along the lateral visual pathway. *Current Biology*, *33*(23), 5035–5047.e8. (Publisher: Elsevier) doi: 10.1016/j.cub.2023.10.015

McMahon, E., & Isik, L. (2023). Seeing social interactions. *Trends in Cognitive Sciences*, *27*(12), 1165–1179. doi: 10.1016/j.tics.2023.09.001

Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends in Cognitive Sciences*, *25*(2), 100–110. doi: 10.1016/j.tics.2020.11.006

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . Rush, A. M. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing.* arXiv. (arXiv:1910.03771 [cs]) doi: 10.48550/arXiv.1910.03771

Wurm, M. F., Caramazza, A., & Lingnau, A. (2017). Action Categories in Lateral Occipitotemporal Cortex Are Organized Along Sociality and Transitivity. *Journal of Neuroscience*, *37*(3), 562–575. doi: 10.1523/JNEUROSCI.1717-16.2016