Investigating the neural computations underlying visual social inference with graph neural networks

Manasi Malik (mmalik16@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
Baltimore, Maryland 21218, United States

Minjae Kim (mkim19@jhu.edu)
Department of Psychological and Brain Sciences, Johns Hopkins University
Baltimore, Maryland 21218, United States

Shari Liu (sliu199@jhu.edu)
Department of Psychological and Brain Sciences, Johns Hopkins University
Baltimore, Maryland 21218, United States

Tianmin Shu (tianmin.shu@jhu.edu)
Department of Computer Science & Department of Cognitive Science, Johns Hopkins University
Baltimore, Maryland 21218, United States

Leyla Isik (lisik@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
Baltimore, Maryland 21218, United States

. .

Abstract

Recognizing people's interactions in visual scenes is a crucial human ability; however, the neural computations that enable this remain largely unknown. Prior work demonstrates that a bottom-up, visual model with relational inductive biases (based on graph-neural-networks) successfully captures human behavior in social interaction judgments, suggesting that relational visual representations may underlie this ability. If relational visual computations are fundamental to social perception, then we should find evidence for them in brain regions that support social perception, such as lateral occipital temporal cortex (LOTC) and posterior STS (pSTS). To test this, we collected fMRI data from adults watching animated shape videos of two agents interacting in a friendly, neutral, or adversarial manner. Preliminary analysis using whole-brain-searchlight representational similarity analysis (RSA) shows a correlation between neural and behavioral representations in both the above social perception regions and the theory of mind network. The graph-neural-network model also explains responses in LOTC and pSTS. In contrast, a matched bottom-up model without relational inductive biases correlates poorly with neural data. Our work suggests brain regions in LOTC and pSTS that support social interaction perception rely on relational visual information, and provides a novel modeling framework for investigating the neural computations underlying social perception and cognition.

Keywords: graph neural networks; social interaction recognition; neuroAI; fMRI; STS

Introduction

Every day we see social interactions between people and extract rich information about their actions and intentions. What computational processes in the human mind and brain enable us to extract such rich information from visual scenes? Previous studies have identified brain regions along the lateral occipital temporal cortex (LOTC) and superior temporal sulcus (STS) involved in social interaction recognition, but to date, there are no computational models that match neural responses in these regions (McMahon & Isik, 2023). Emerging behavioral and neural evidence supports a bottom-up theory, suggesting that social interaction recognition is primarily a visual process, distinct from mental-state inference (Abassi & Papeo, 2020; Isik et al., 2017; Masson & Isik, 2021; Papeo et al., 2017; Hafri et al., 2018; Su et al., 2016; Vestner et al., 2019; Gandolfo et al., 2024; McMahon & Isik, 2023). Previously, this hypothesis was dismissed due to the poor performance of bottom-up, visual models in matching human judgments of social interactions (Ullman et al., 2009; Isik et al., 2020). However, a recent study showed that a bottom-up model, equipped with relational inductive biases (based on a graph-neural-network, called SocialGNN), can match people's judgments about social interactions (e.g., whether an interaction is "friendly", "neutral", or "adversarial") and performs significantly better than a matched model with the same input and training, but without the relational inductive bias (VisualRNN) (Malik & Isik, 2023). These results suggest that relational information may be critical to human social interaction representations.

Here we ask whether the brain regions representing social interactions rely on relational visual representations as captured by SocialGNN. To achieve this, we employed representational similarity analysis (RSA) to compare neural data with behavior and two computational models of social interaction recognition: SocialGNN (Fig. 1b), the bottom-up graph-neural-network based model (Malik & Isik, 2023), and VisualRNN (Fig. 1c), a control model that shares the same broad architecture and input as SocialGNN but lacks the graph structure and processing (Malik & Isik, 2023).

Methods

fMRI Experiment

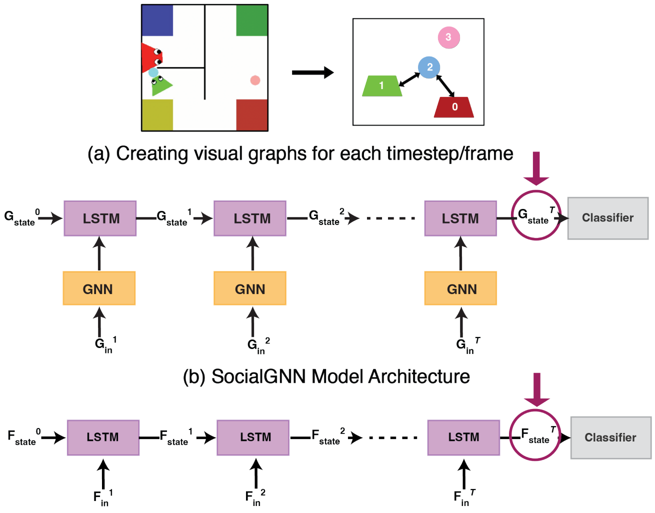
Participants This study received ethical approval from the Johns Hopkins School of Medicine Institutional Review Board. fMRI data was collected from four participants who provided written informed consent before the experiment and were monetarily compensated.

Stimuli Our stimulus set consisted of 50 videos from the PHASE dataset (Netanyahu et al., 2021), which includes 500 videos of two agents moving around in a simple 2D environment, navigating around obstacles and manipulating objects, designed to resemble real-life social interactions. Each video was generated via a physical simulator and hierarchical planner, based on specified goals of the agents, their relative strengths, and the objects in the scene. For more details on the dataset see Netanyahu et al. (2021). We selected 50 videos, and trimmed each to keep the middle 10 seconds.

Experiment Paradigm and Data Preprocessing Following an anatomical scan, participants watched each of the 50 videos, five times across separate runs. The video presentation order in every run was shuffled. Data preprocessing was done using fMRIprep 21.0.2 (Esteban et al., 2019), and we then used GLMsingle (Prince et al., 2022) to estimate per-voxel, per-video BOLD responses.

Computational Models and Behavioural Data

SocialGNN SocialGNN is a graph-neural-network model that incorporates relational inductive biases to recognize social interactions from visual scenes (Fig. 1b) (Malik & Isik, 2023). For each video, it takes in a graph representation (Fig. 1a, right) for each frame (Fig. 1a, left). Specifically, the nodes in these graphs represent agents/objects in the scene, node features are visuospatial features of each entity (such as current position, velocity, etc.), and binary edges represent physical contact between the entities. At each timestep, SocialGNN processes new input graphs (GNN module), and combines it with the learned representations from prior timesteps (LSTM module). Representation at the final time-step is passed



(a) Creating visual graphs for each timestep/frame

(b) SocialGNN Model Architecture

(c) VisualRNN Model Architecture

Figure 1: Computational Models Architectures and Representations. Output from the last RNN step (circled in red) is taken as the model's representation of that video

through a linear classifier to predict the type of social interaction ("friendly", "neutral", or "adversarial"). For our analysis, SocialGNN was trained on 400 videos, and predictions were made on the 50 held-out videos used in the fMRI experiment. We took the output from the last RNN step as the model representation for each video to create the representational dissimilarity matrix (RDM) (Fig. 1b).

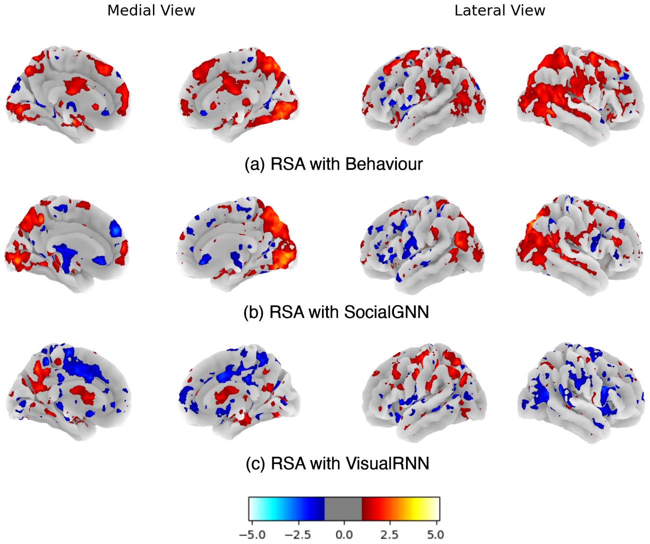
VisualRNN VisualRNN has the same broad RNN architecture and input information as SocialGNN but lacks the graph structure and graph processing (Fig. 1c) (Malik & Isik, 2023). Essentially, the node features used in the visual graphs for SocialGNN input, are instead concatenated for all entities in the scene and directly input to the LSTM. Like SocialGNN, we made predictions using this model on the 50 held-out videos used in the fMRI experiment, and then created an RDM using the output from the last RNN step.

Behaviour To compare fMRI responses to human judgments on the PHASE dataset, we utilized human judgments from a prior study where participants rated the relationship in each video as 'friendly', 'neutral', or 'adversarial' (Malik & Isik, 2023). There were at least ten ratings per video and we used the normalized counts of these as the representation to create the human behavior RDM.

Results

To assess the representational similarity between various brain regions and computational models, as well as behavioral judgments, we conducted a whole-brain searchlight RSA asking where the representational geometry of brain activity, evoked by the 50 videos, is explained by (i) human behavioral judgments and (ii) model representations of the those videos. We used the pattern of estimated BOLD responses for each video as the representation to create neural RDMs. All RDMs (neural, model, behaviour) were based on Pearson's correlation distance.

There were two main findings. First, we found a correlation between neural RDMs and the behavioral RDM within the visual cortex, lateral occipital temporal cortex (LOTC), superior temporal sulcus (STS), temporoparietal junction (TPJ), and medial prefrontal cortex (mPFC) (Fig. 2a). This similarity confirms prior work showing that these regions play a critical role in inferring social interactions from visual scenes (McMahon & Isik, 2023). Second, SocialGNN correlated with many of the regions showing a match to behavior, including the LOTC and posterior STS (Fig. 2b). In contrast, despite its similar input and training, VisualRNN did not exhibit any correlations with neural representations in these brain areas (Fig. 2c).



(a) RSA with Behaviour

(b) RSA with SocialGNN

(c) RSA with VisualRNN

Figure 2: Whole-brain searchlights RSA (Group-Level Z-Maps).

Conclusion

Our preliminary findings reveal a representational match between human social interaction judgements and regions previously implicated in both social interaction perception (including LOTC and the STS) and theory of mind (including mPFC and TPJ), replicating prior findings. We further find that SocialGNN, a bottom-up model with relational inductive biases, matches neural representations in LOTC and the STS significantly better than a matched visual model. These results provide the first computational modeling account of neural activity in these brain areas, and preliminary evidence that these regions may be structuring visual information relationally to facilitate social interaction recognition. In ongoing work, we are validating these findings with more experimental subjects. We are also exploring additional models to understand how the human brain combines visual processing and mental state inference to recognize social scenes.

Acknowledgments

This work was funded by NIMH R01MH132826 awarded to L.I.

References

Abassi, E., & Papeo, L. (2020). The representation of two-body shapes in the human visual cortex. *Journal of Neuroscience*, *40*(4), 852–863. doi: 10.1523/JNEUROSCI.1378-19.2019

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., . . . others (2019). fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, *16*(1), 111–116.

Gandolfo, M., Abassi, E., Balgova, E., Downing, P. E., Papeo, L., & Koldewyn, K. (2024). Converging evidence that left extrastriate body area supports visual sensitivity to social interactions. *Current Biology*, *34*(2), 343–351.

Hafri, A., Trueswell, J. C., & Strickland, B. (2018). Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. *Cognition*, *175*, 36–52.

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *114*(43), E9145–E9152.

Isik, L., Mynick, A., Pantazis, D., & Kanwisher, N. (2020). The speed of human social interaction perception. *NeuroImage*, *215*, 116844.

Malik, M., & Isik, L. (2023). Relational visual representations underlie human social interaction recognition. *Nature Communications*, *14*(1), 7317.

Masson, H. L., & Isik, L. (2021). Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *NeuroImage*, *245*, 118741.

McMahon, E., & Isik, L. (2023). Seeing social interactions. *Trends in Cognitive Sciences*.

Netanyahu, A., Shu, T., Katz, B., Barbu, A., & Tenenbaum, J. B. (2021). Phase: Physically-grounded abstract social events for machine social perception. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 845–853).

Papeo, L., Stein, T., & Soto-Faraco, S. (2017). The two-body inversion effect. *Psychological science*, *28*(3), 369–379.

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022). Improving the accuracy of single-trial fmri response estimates using glmsingle. *Elife*, *11*, e77599.

Su, J., Van Boxtel, J. J., & Lu, H. (2016). Social interactions receive priority to conscious perception. *PloS one*, *11*(8), e0160468.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, *22*.

Vestner, T., Tipper, S. P., Hartley, T., Over, H., & Rueschemeyer, S.-A. (2019). Bound together: Social binding leads to faster processing, spatial distortion, and enhanced memory of interacting partners. *Journal of Experimental Psychology: General*, *148*(7), 1251.